*COURSERA CAPSTONE PROJECT*

# ANALYSIS OF PROPERTY INVESTMENT IN LONDON AREA BASED ON VARIOUS AMENITIES AND PRICE FACTORS

By Hemendrakumar Jasani

OCTOBER 29, 2020
IBM APPLIED DATA SCIENCE
Coursera course

# Capstone Project - The Battle of Neighbourhoods.

## Research Question section

### Background

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. More specifically, four overlooked cracks suggest that the London market may be in worse shape than many realize: hidden price falls, record-low sales, homebuilder exodus and tax hikes addressing overseas buyers of homes in England and Wales.

### Research Question

In this scenario, it is urgent to adopt machine learning tools in order to assist homebuyer 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 clientele in to purchase a suitable real estate in London in this uncertain economic and financial scenario?

To solve this business problem, we are going to cluster London neighbourhoods in order to recommend venues and the current average price of real estate where homebuyers can make a real estate investment. We will recommend profitable venues according to amenities and essential facilities surrounding such venues i.e. elementary schools, high schools, hospitals & grocery stores.

## Data section

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.

# Methodology section

The Methodology section will describe the main components of our analysis and predication system. The Methodology section comprises four stages:

1. Collect Inspection Data
2. Explore and Understand Data
3. Data preparation and pre-processing
4. Modelling

### 1. Collect Inspection Data and Explore and Understand Data

After importing the necessary libraries, we download the data from the HM Land Registry website. We read the dataset that we collected from the HM Land Registry website into a pandas' data frame and display the first five rows of it as follows:

In [160]: `df_HousePrice.head()`

Out[160]:

| | {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 | GATESHEAD | TYNE AND WEAR | B | A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | SUNDERLAND | TYNE AND WEAR | A | A |
| 1 | {79A74E22-41E4-1289-E053-6B04A8C01627} | 164950 | 2018-06-29 00:00 | SR2 0FD | S | Y | F | 6 | NaN | WILSHIRE CLOSE | NaN | SUNDERLAND | SUNDERLAND | TYNE AND WEAR | A | A |
| 2 | {79A74E22-41E5-1289-E053-6B04A8C01627} | 224950 | 2018-06-29 00:00 | SR2 0FA | D | Y | F | 47 | NaN | WOODHAM DRIVE | NaN | SUNDERLAND | SUNDERLAND | TYNE AND WEAR | A | A |
| 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 | SUNDERLAND | TYNE AND WEAR | A | A |
| 4 | {79A74E22-41E7-1289-E053-6B04A8C01627} | 144395 | 2018-02-23 00:00 | NE31 2EL | T | Y | F | 9 | NaN | TURNBERRY DRIVE | NaN | HEBBURN | SOUTH TYNESIDE | TYNE AND WEAR | A | A |

In [161]: `df_HousePrice.shape`

Out[161]: `(1031509, 16)`

*Figure 1 Raw data.*

Our dataset consists of over 1031509 rows and 16 columns. We will now prepare and pre-process data accordingly. Before using data, we will have to explore and understand it.

## 3. Data preparation and pre-processing

At this stage, we prepare our dataset for the modelling process, opting for the most suitable machine learning algorithm for our scope. Accordingly, we perform the following steps:

1. Rename the column names
2. Format the date column
3. Sort data by date of sale
4. Select data only for the city of London
5. Make a list of street names in London
6. Calculate the streetwise average price of the property
7. Read the streetwise coordinates into a data frame, eliminating recurring word London from individual names
8. Join the data to find the coordinates of locations which fit into client's budget
9. Plot recommended locations on London map along with current market prices

Result of above processing is as below,

```
In [172]: df_HousePrice_affordable1 = df_HousePrice_affordable.drop(columns=['city_coord'])
          df_HousePrice_affordable1
```

Out[172]:

| | 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 | 51.499890 | -0.189103 |
| **422** | APPLEGARTH ROAD | 2400000.0 | 53.749244 | -0.326780 |
| **857** | BARONSMEAD ROAD | 2375000.0 | 51.477315 | -0.239457 |
| **...** | ... | ... | ... | ... |
| **13733** | WILFRED STREET | 2410538.5 | 42.508350 | -82.921795 |
| **13759** | WILLOW BRIDGE ROAD | 2425000.0 | 53.640874 | -1.200462 |
| **13779** | WILSON STREET | 2257500.0 | 45.700486 | -121.521875 |
| **13808** | WINCHENDON ROAD | 2350000.0 | 42.645331 | -71.954714 |
| **13845** | WINGATE ROAD | 2206400.0 | 41.043293 | -80.621035 |

159 rows × 4 columns

*Figure 2 Data after getting location coordinates.*

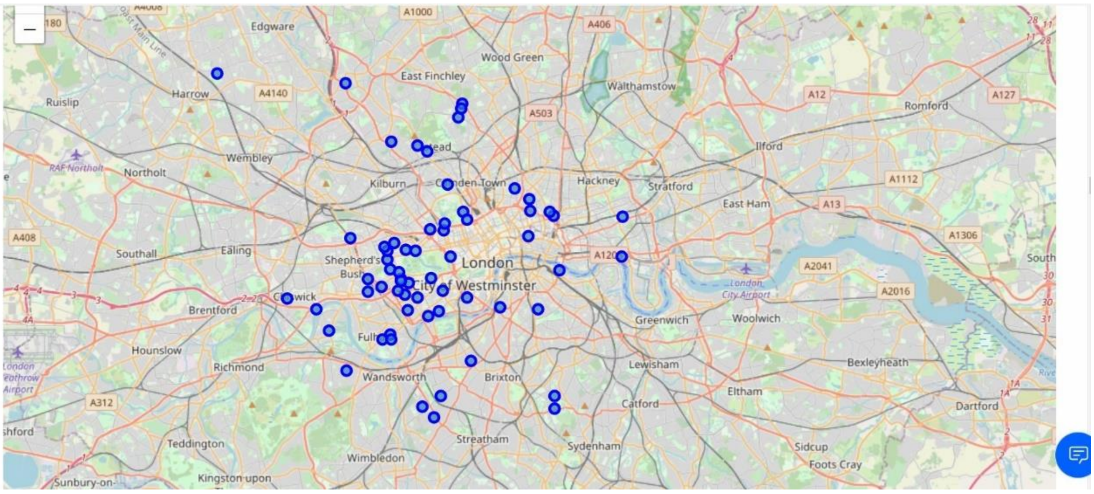


*Figure 3 primary map with all marked locations without clustering*

In [189]: venues_sorted.head()

Out[189]:

| | 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 | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | ALBION SQUARE | Café | Indian Restaurant | Restaurant | Coffee Shop | Pub | Bar | Park | Burger Joint | Museum | Art Gallery |
| 1 | ANHALT ROAD | Hotel | Movie Theater | Gym | Intersection | Coffee Shop | Factory | Electronics Store | English Restaurant | Escape Room | Ethiopian Restaurant |
| 2 | ANSDELL TERRACE | Juice Bar | Restaurant | Hotel | Clothing Store | Italian Restaurant | Pub | Indian Restaurant | Café | Sporting Goods Shop | French Restaurant |
| 3 | APPLEGARTH ROAD | Bar | Sandwich Place | Auto Dealership | Nightclub | Casino | Fabric Shop | Electronics Store | English Restaurant | Escape Room | Ethiopian Restaurant |
| 4 | BARONSMEAD ROAD | Food & Drink Shop | Pizza Place | Indie Movie Theater | Restaurant | Nature Preserve | Breakfast Spot | Coffee Shop | Sports Club | Thai Restaurant | Farmers Market |

*Figure 4 Dataframe containing streetwise most common venues availability.*

london_grouped_clustering.head(30)

Out[197]:

| | 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 Mo Comm Ven |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 196 | ALBION SQUARE | 2.450000e+06 | -41.273758 | 173.289393 | 2 | Café | Indian Restaurant | Restaurant | Coffee Shop | Pub | Bar | Park | Burger Joint | Muse |
| 390 | ANHALT ROAD | 2.435000e+06 | 29.712770 | -98.094806 | 0 | Hotel | Movie Theater | Gym | Intersection | Coffee Shop | Factory | Electronics Store | English Restaurant | Esca Ro |
| 405 | ANSDELL TERRACE | 2.250000e+06 | 51.499890 | -0.189103 | 1 | Juice Bar | Restaurant | Hotel | Clothing Store | Italian Restaurant | Pub | Indian Restaurant | Café | Sport Goods Sh |
| 422 | APPLEGARTH ROAD | 2.400000e+06 | 53.749244 | -0.326780 | 0 | Bar | Sandwich Place | Auto Dealership | Nightclub | Casino | Fabric Shop | Electronics Store | English Restaurant | Esca Ro |
| 857 | BARONSMEAD ROAD | 2.375000e+06 | 51.477315 | -0.239457 | 3 | Food & Drink Shop | Pizza Place | Indie Movie Theater | Restaurant | Nature Preserve | Breakfast Spot | Coffee Shop | Sports Club | T Restaur |
| 983 | BEAUCLERC ROAD | 2.480000e+06 | 30.211452 | -81.617981 | 2 | Pizza Place | Spa | Automotive Shop | Harbor / Marina | Factory | Electronics Store | English Restaurant | Escape Room | Ethiop Restaur |
| 1105 | BELVEDERE DRIVE | 2.340000e+06 | 38.072439 | -78.459970 | 3 | Pool | Playground | Athletics & Sports | Factory | Egyptian Restaurant | Electronics Store | English Restaurant | Escape Room | Ethiop Restaur |
| 1218 | BICKENHALL STREET | 2.208500e+06 | 51.521201 | -0.158908 | 1 | Coffee Shop | Italian Restaurant | Hotel | Café | Gastropub | Pizza Place | Chinese Restaurant | Bar | Bak |
| 1256 | BIRCHLANDS AVENUE | 2.217000e+06 | 51.448394 | -0.160468 | 1 | Pub | Lake | Coffee Shop | French Restaurant | Bakery | Chinese Restaurant | Train Station | Brewery | Pizza Pl |
| 1556 | BRAMPTON GROVE | 2.456875e+06 | 51.589961 | -0.318525 | 2 | Home Service | Zoo | Factory | Egyptian Restaurant | Electronics Store | English Restaurant | Escape Room | Ethiopian Restaurant | Event Spa |

*Figure 5 Final Data frame which contains street name as index with data of price, coordinates and most common venues of that street.*

We can now proceed to the Modelling phase. We will analyse neighbourhoods to recommend real estates where home buyers can make a real estate investment. We will then recommend profitable venues according to amenities and essential facilities surrounding such venues i.e. elementary schools, high schools, hospitals & grocery stores.

## 4. Modeling

After exploring the dataset and gaining insights into it, we are ready to use the clustering methodology to analyze real estates. We will use 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.

The result of K-clustering shows that we can recognise the London neighbourhoods based on different 5 clusters based on prise value and available most common venue (e.g. café, park, gym etc.) This can be utilised while deciding investment in real estate market.

Based on clustering below clusters shows predictive model,

In [199]:
```
london_grouped_clustering.loc[london_grouped_clustering['Cluster Labels'] == 0, london_grouped_clustering.columns[[1] + list(range(
```

Out[199]:

| | Avg_Price | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 390 | 2435000.0 | Hotel | Movie Theater | Gym | Intersection | Coffee Shop | Factory | Electronics Store | English Restaurant | Escape Room | Ethiopian Restaurant |
| 422 | 2400000.0 | Bar | Sandwich Place | Auto Dealership | Nightclub | Casino | Fabric Shop | Electronics Store | English Restaurant | Escape Room | Ethiopian Restaurant |
| 1635 | 2397132.0 | Convenience Store | Coffee Shop | Breakfast Spot | Grocery Store | Park | Coworking Space | Falafel Restaurant | English Restaurant | Escape Room | Ethiopian Restaurant |
| 1800 | 2400000.0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 2162 | 2425000.0 | Clothing Store | Electronics Store | Chinese Restaurant | Gym | Department Store | Basketball Court | Stationery Store | American Restaurant | Convenience Store | Women's Store |

click to expand output; double click to hide output

*Figure 6 Cluster 0*

In [200]:
```
london_grouped_clustering.loc[london_grouped_clustering['Cluster Labels'] == 1, london_grouped_clustering.columns[[1] + list(range(
```

Out[200]:

| | Avg_Price | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 405 | 2250000.0 | Juice Bar | Restaurant | Hotel | Clothing Store | Italian Restaurant | Pub | Indian Restaurant | Café | Sporting Goods Shop | French Restaurant |
| 1218 | 2208500.0 | Coffee Shop | Italian Restaurant | Hotel | Café | Gastropub | Pizza Place | Chinese Restaurant | Bar | Bakery | Pub |
| 1256 | 2217000.0 | Pub | Lake | Coffee Shop | French Restaurant | Bakery | Chinese Restaurant | Train Station | Brewery | Pizza Place | Cricket Ground |
| 2229 | 2200000.0 | Trail | Zoo | Dry Cleaner | Egyptian Restaurant | Electronics Store | English Restaurant | Escape Room | Ethiopian Restaurant | Event Space | Exhibit |
| 2642 | 2250000.0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

In [201]:
```
london_grouped_clustering.loc[london_grouped_clustering['Cluster Labels'] == 2, london_grouped_clustering.columns[[1] + list(range(
```

*Figure 7 Cluster 1*

```
In [201]: london_grouped_clustering.loc[london_grouped_clustering['Cluster Labels'] == 2, london_grouped_clustering.columns[[1] + list(range(
```

Out[201]:

| | Avg_Price | 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** | 2450000.0 | Café | Indian Restaurant | Restaurant | Coffee Shop | Pub | Bar | Park | Burger Joint | Museum | Art Gallery |
| **983** | 2480000.0 | Pizza Place | Spa | Automotive Shop | Harbor / Marina | Factory | Electronics Store | English Restaurant | Escape Room | Ethiopian Restaurant | Event Space |
| **1556** | 2456875.0 | Home Service | Zoo | Factory | Egyptian Restaurant | Electronics Store | English Restaurant | Escape Room | Ethiopian Restaurant | Event Space | Exhibit |
| **1983** | 2492500.0 | Supermarket | English Restaurant | Park | Gym | Dry Cleaner | Hardware Store | Rental Car Location | Coffee Shop | Discount Store | American Restaurant |
| **2139** | 2461000.0 | Pub | Factory | Egyptian Restaurant | Electronics Store | English Restaurant | Escape Room | Ethiopian Restaurant | Event Space | Exhibit | Fabric Shop |

*Figure 8 Cluster 2*

```
In [202]: london_grouped_clustering.loc[london_grouped_clustering['Cluster Labels'] == 3, london_grouped_clustering.columns[[1] + list(range(
```

Out[202]:

| | Avg_Price | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **857** | 2375000.0 | Food & Drink Shop | Pizza Place | Indie Movie Theater | Restaurant | Nature Preserve | Breakfast Spot | Coffee Shop | Sports Club | Thai Restaurant | Farmers Market |
| **1105** | 2340000.0 | Pool | Playground | Athletics & Sports | Factory | Egyptian Restaurant | Electronics Store | English Restaurant | Escape Room | Ethiopian Restaurant | Event Space |
| **2071** | 2375000.0 | Pub | Park | Indian Restaurant | Pizza Place | Hotel | Yoga Studio | Breakfast Spot | Juice Bar | Outdoor Sculpture | Sushi Restaurant |
| **2132** | 2379652.7 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **2948** | 2367500.0 | Hotel | Pub | Garden | Café | Coffee Shop | Chinese Restaurant | Bar | Cocktail Bar | Italian Restaurant | Mediterranean Restaurant |

*Figure 9 Cluster 3*

```
In [203]: london_grouped_clustering.loc[london_grouped_clustering['Cluster Labels'] == 4, london_grouped_clustering.columns[[1] + list(range(
```

Out[203]:

| | Avg_Price | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **2246** | 2.300000e+06 | Farm | Falafel Restaurant | Electronics Store | English Restaurant | Escape Room | Ethiopian Restaurant | Event Space | Exhibit | Fabric Shop | Factory |
| **2410** | 2.286679e+06 | Café | Pub | Italian Restaurant | Bar | Coffee Shop | Convenience Store | Park | French Restaurant | Supermarket | Portuguese Restaurant |
| **2690** | 2.287500e+06 | Pub | Reservoir | Gift Shop | Harbor / Marina | Art Museum | Brewery | Coworking Space | Factory | Electronics Store | English Restaurant |
| **3381** | 2.298000e+06 | Hotel | Zoo | Factory | Egyptian Restaurant | Electronics Store | English Restaurant | Escape Room | Ethiopian Restaurant | Event Space | Exhibit |
| **4289** | 2.265000e+06 | Pub | Factory | Egyptian Restaurant | Electronics Store | English Restaurant | Escape Room | Ethiopian Restaurant | Event Space | Exhibit | Fabric Shop |

*Figure 10 Cluster 4*

# Discussion section

We may discuss our results under two main perspectives.

First, we may examine them according to neighborhoods/London areas. It is interesting to note that, 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 may analyze our results according to the five clusters we have produced. Even though, all clusters could praise an optimal range of facilities and amenities, we have found two main patterns. The first pattern we are referring to, i.e. Clusters 0, 2 and 4, may target home buyers prone to live in 'green' areas with parks, waterfronts. Instead, the second pattern we are referring to, i.e. Clusters 1 and 3, may target individuals who love pubs, theatres and soccer.

# 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 London areas. although West London (Notting Hill, Kensington, Chelsea, Marylebone) and North-West London 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 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 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.