Analysis of London Real Estate Market

Recommending Similar Properties on Sale

May 4th, 2020

---

2. Methodology and Data

I am interested in implementing an unsupervised method to avoid labeling input by the agents, which would introduce subjectivity to the evaluations. Given the nature of the problem requires identifying similarity across different data points, I am planning on using K-means clustering to identify properties similar to each other.

For this project, I will particularly focus on London real estate, which is one of the prime locations throughout the world. Given the cost of housing and extensive availability of properties throughout the year, making relevant and accurate recommendations to potential buyers has a lot of value for real estate businesses in London. I also choose to focus on properties on sale as opposed to rental units.

Clustering similar properties on sale in London real estate market requires combining several datasets:

- List of properties on sale
  In order to get up-to-date data on property listings in London, I will use a web scraping app to download the data from RightMove.co.uk. The app used is called rightmove-webscraper and is available on GitHub by toby-p. This app is embedded in my code.
  While downloading the data from RightMove.co.uk, I particularly focused on properties on sale, the listings from the past 14 days and those with

listing price less than or equal to £2 million. Among other information, for each property, this dataset includes:

  - Address
  - Price
  - Number of bedrooms

- List of postcodes and district names in London
  The name of the district that each property is located in is not available in the RightMove dataset. Therefore, I will use web scraping to download the table that includes the list of postcodes and district names in London from https://en.wikipedia.org/wiki/London_postal_district

- Coordinates of each propery using OpenCage API
  RightMove dataset provides the address, but not the exact coordinates of the listed properties. In order to obtain this information, I used OpenCage API to convert address information to coordinates. This way, for each listing, I obtained
  - Latitude, and
  - Longitude,

  which will be inputs to the queries using Foursquare API.

- List of venues nearby for each property using Foursquare API
  The last component of the data is the list of venues in 500 meters proximity of each listing obtained through Foursquare API using the latitude and longitude information. In order to simplify the venue categories and reduce the computational effort, I will particularly focus on the top 20 venue categories around each property.

  This last component adds 20 more columns to the dataset that includes the number of pubs, coffee shops, cafés, hotels, grocery stores, Italian restaurants, pizza places, parks, gyms / fitness centers, sandwich places, bakeries, bars, Indian restaurants, burger joints, supermarkets, bus stops, cocktail bars, plazas and theaters.

Once the aforementioned datasets are merged, we will have both the data describing the property such as price, number of bedrooms, district, and data describing the neighborhood such as the number of top 20 most common venues in 500m. vicinity.

Using this dataset, my goals are two-fold:

- First, analyze this data to understand the cost and availability of properties on sale in London and the relative expense of different districts.

- Second, apply K-means clustering on this dataset to identify similar clusters of properties. In practice, the output of this analysis would be the basis for a realtor's recommendation to their customers. For example, if a customer got in touch with the realtor regarding a specific property in Cluster 1, even if the customer is not interested in buying this property, the realtor can recommend other properties in Cluster 1 as they are similar across multiple dimensions.