Determining suitable neighbourhoods to buy property in London

Martin Manjolo

April 19, 2020

Introduction

Background

London is one of the world's largest tech centres and a sought-after destination for millennials from all over the world looking to make a mark. A city as large and as modern as London offers several opportunities for growth and is renowned for its high standard of life. This also makes a challenge for prospective immigrants to find a home they can purchase and start a life in.

Problem

In this project we will try to find the optimal neighbourhood to buy a home. We will focus on potential homeowners looking to buy a home in London, England. One of the most important considerations is the proximity to Central London, as well as affordability.

Interest

The intended audience are young millennials moving to the UK with a moderate-income base, looking to buy a 2 bedroomed flat in or around the Greater London Area. We will do this by creating clusters of neighbourhoods in London based on selected amenities and average real estate prices.

Data acquisition and cleaning

Data sources

Data on London properties and relative price data is extracted from the HM Land registry available here:

https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads

We employ the price paid data for the year 2020, last updated on 27 March 2020.

Additionally, we employ the use of geographical coordinate data provided by FreeMapTools. This proved useful as we faced challenges geocoding the coordinates. UK postal code and geographical data is available here:

https://www.freemaptools.com/download-uk-postcode-lat-lng.htm

We also employ Foursquare data to discover recommended amenities and venues in the neighbourhoods in order to select and recommend idea neighbourhoods for considerations. Based on our problem, factors that will influence our decision are:

- We are looking at a 2 bedroomed apartment or flat as close as possible to Central London
- The price range we are working with is between 500,000 GBP and 510,000 GBP
- Venues and essential amenities should be located within 500 metres of the property.

Data cleaning

Data downloaded from the HM Land Registry and FreeMapTools were combined into one table. I decided to only use data for 2020 as it provides the most recent information on prices paid for various properties in the UK. Fortunately, limited cleaning was required as the data sources are well formatted and almost all the data fields required are available for our purposes.

The price paid data lacked meaningful column names, so I created a list and added it to the data frame based on the following column titles as indicated on the HM Land Registry website:

- **TUID** – The unique identifier of the property
- **Price** – The actual price paid for the property
- **Date_Transfer** - The date and time the transaction was completed and recorded into the registry
- **Postcode** – The post code of the neighbourhood where the property is located
- **Prop_Type** – A classification of the type of property.
- **Old_New** – An indication of whether the property is newly built or an older property
- **Duration** – An indication of the type of ownership type attached to the property
- **PAON** – The 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** – The name of the street where the property is located
- **Locality** – The neighbourhood in which the property is located
- **Town_City** – The town or city where the property is located

- **District** – The district where the property is located
- **County** – The county where the property is located
- **PPD_Cat_Type** – The price paid data category type
- **Record_Status** – The record status in the HM registry

Feature selection

After data merging, there where 70,376 samples and 19 features in the data. Upon examining the features, it was clear that there was some redundancy in the features. For our purposes, we only required the street names, prices, post code and geographical coordinates. We discarded the rest of the features and set the average price as our dependent variable.

Exploratory Data Analysis

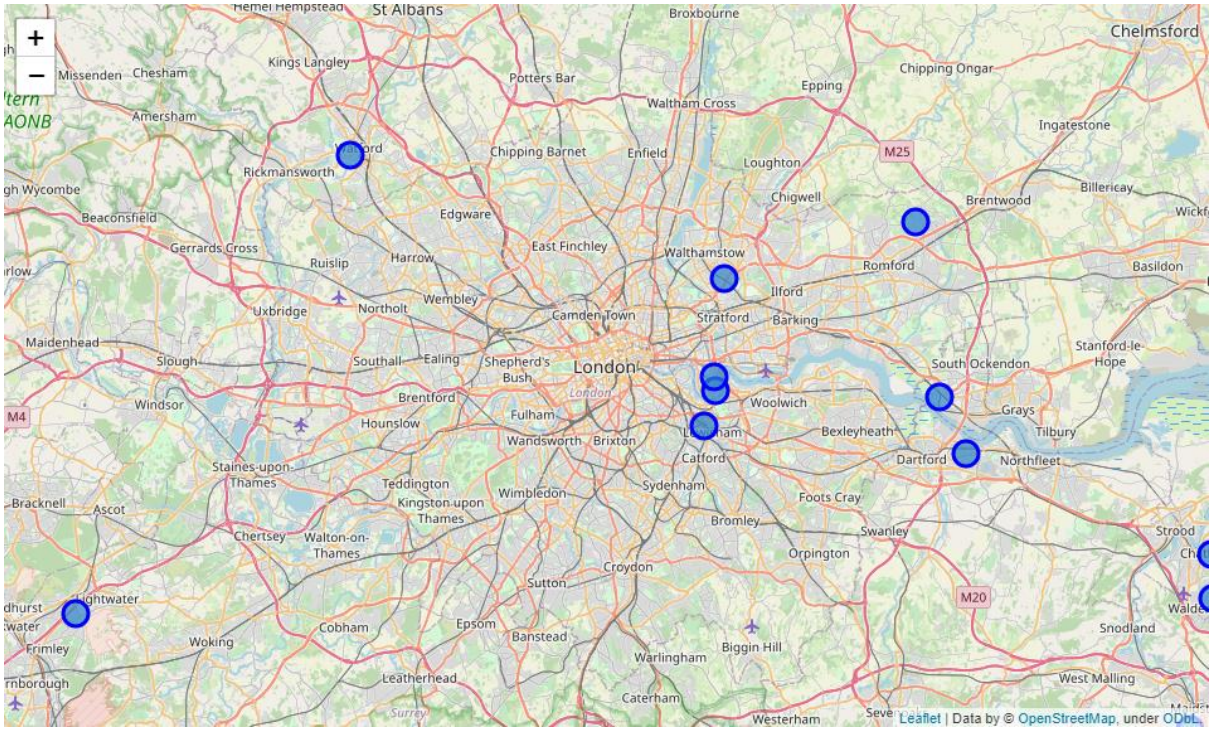
I chose to use the average price as the target variable. The average price was calculated from the average paid prices in the same neighbourhood along the same street. I arranged the samples in ascending order of date of transfer, and then filtered the data to only include properties in the Greater London county. Once done, I created a dataframe that filtered the average prices obtained by the limits we had set, a property value between 500,000 GBP and 510,000 GBP. This returned 123 samples. After this, I re-added the geographical coordinates that would allow us to visualize the clusters that would be developed in the modelling.

Modelling

Visualization

Once all the data had been collected, I visualized the data to get a sense of how spread out our target neighbourhoods where. This involved obtaining the geographical coordinates of London, and them creating a map with the selected neighbourhoods as markers. The following map resulted from this visualization:

Map of London with selected neighbourhoods



Initial indications are that properties in this price bracket are rare and few. This could be attributed to the high cost of properties in and around London, making this segment of the market highly competitive.

Foursquare API

We employed the foursquare API to get information on venues around our target neighbourhoods based on amenities and essential facilities such as schools, hospitals and grocery stores. We limited our range to only include the top 100 venues within 500 metres of our neighbourhoods. We analysed the data and grouped the venues by neighbourhood and by taking the mean of the frequency of occurrence of each category. We then selected the top 10 most common venues in each neighbourhood.

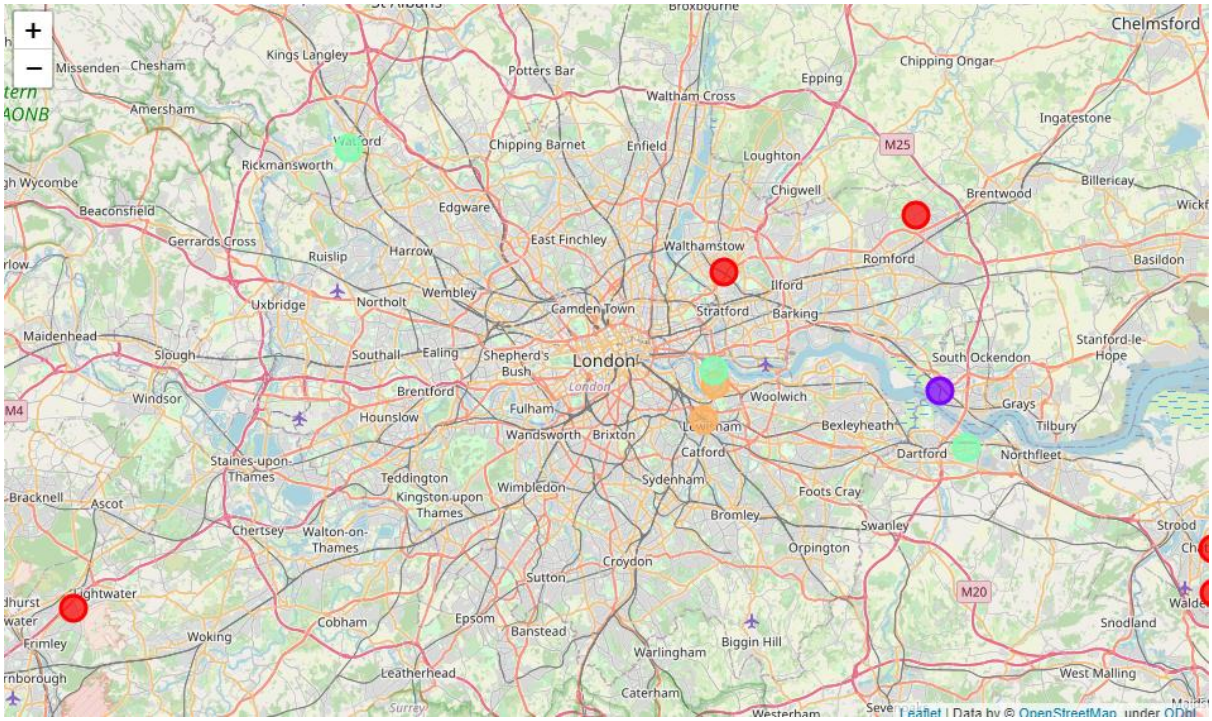
K-Means Clustering

There are many models for clustering out there. For this project, we settled on K-Means clustering. K-means is vastly used for clustering in many data science applications, especially useful to quickly discover insights from unlabelled data. Some real-world applications of k-means include:

- Customer segmentation
- Understand what the visitors of a website are trying to accomplish
- Pattern recognition
- Machine learning
- Data compression

Based on the data we obtained from Foursquare, we clustered the neighbourhoods into 5 clusters. We created a dataframe that included the cluster as well as the top 10 venues for each neighbourhood. We then visualized the neighbourhoods. The results of are shown below:

Map of London with clustered neighbourhoods



Results and Discussion

Our analysis shows that at our price point, there are more options for homes further outside of Central London. Fewer clusters closer to Central London are indicative of the fact this segment of the market is highly competitive and that properties rarely become available for sale.

Our analysis is based on looking at neighbourhoods that are within the 500K – 510K GBP price range, for a 2 bedroomed flat or apartment near Central London. By analysing price paid data in the region, we narrowed down a sample set of over 70K properties to 123 properties that had been purchased in 2020. Focusing on the neighbourhoods where these properties where located, we cross referencing with Foursquare data on venues and amenities within a 500-metre radius of each identified location, we were able to identify the most suitable neighbourhoods with the right balance of amenities and venues.

Candidate neighbourhoods where clustered to create zones of interest with the greatest number of candidates. As a result, the following streets and neighbourhoods are our top picks, offering the best combination of price, location and amenities:

- London
- Manchester
- Wallington
- South Croydon

Conclusion

The project identified neighbourhoods near Central London that properties within our set price budget and had the right blend of amenities and venues. We have identified 4 neighbourhoods that meet the basic criteria. The final decision on the right neighbourhood will be made by prospective homeowners, considering the characteristics of the neighbourhoods, proximity to favourite venues and amenities, and ease of travel to and from Central London.