Predicting house price differences between London Boroughs

Alec Wilson - 12/1/2020

Contents:

1.1 Background

House prices in London are amongst the highest in the world, and, are the highest in the UK compared to other UK regions. Within London, however, there is also a large variation in the average price levels between the 32 London boroughs. Understanding the determinants of these differences is important as it enables firms and households to understand both how to price and value properties for future house building and when planning where to buy. An example of a determinant of house prices may be whether there are good schools nearby or strong transport links.

1.2 Problem

The data which will help solve the problem of determining house prices in various London boroughs will be descriptive data about the individual boroughs, e.g. the crime rate and number of outstanding schools, and, data which helps describe the quality of life in a particular borough.

1.3 Interest

A number of parties would be interested in the determinants for house prices between boroughs. These include commercial interests from sectors such as estate agents and house builders, who, once the determinants are understood, can use the insights to predict future trends.

2.1 Data Sources

The core data for house prices, crime levels, and population were sourced from 3 Kaggle data sets, link. Further data regarding the number of Ofsted (UK schools regulator) rated ‘outstanding’ primary schools was found from here and a commutability score, which assesses various factors including average cost to commute and average time commuting each way, was sourced from here. Data used for exploratory data analysis was also sourced using the Foursquare API. The co-ordinates for the Boroughs needed to be scraped off Wikipedia.

2.2 Data Cleaning

Initially the datasets were imported from .csv files into the IBM Watson Studio environment. The data regarding the total crimes in each London borough was given for each month since late 2016. I decided to focus on the last 12 months of available data and then aggregated the last 12 months of crimes for each borough. The categories of crime were removed from the data frame, further analysis of the types of crime present would be interesting but were not the focus of this analysis. A crime rate per 1000 people in each borough was then created using the population dataset mentioned below.

The data for house prices included prices since January 1995. I was only interested in the latest 12 months available to begin cross-sectional analysis. The data needed to be transposed as the rows were made up of the dates and the columns individual boroughs. In order to merge the data frames together the boroughs were moved to an individual row and indexed. An average of the last 12 months of house prices was taken.

The population data columns were renamed to 'borough' and 'population' and the boroughs were indexed.

The co-ordinates for each borough were scraped off Wikipedia using Beautiful Soup and the latitudes and longitudes were decimalised in order for them to be used by the Foursquare API.

The data for number of outstanding schools and commutability score were found through the links above and extracted into a .csv file. The .csv file was then uploaded into the environment, the .csv file was already formatted in a consistent way. A 'good school' rate per 100,000 people was then created by dividing the number of outstanding schools by the population.

All of this data was then merged into one data frame which was now ready for data analysis.

2.3 Feature Selection

The dependent variable of this problem is the average house price of the individual boroughs.

Whilst cleaning the data a number of variables were created which were normalised for the population size of the individual boroughs. The variables for total crimes, number of outstanding schools and population were therefore not used when analysing the data.

The Foursquare API was used in the exploratory data analysis of the problem. An example of this was querying all of the train and underground stations in London and mapping these using Folium in order to learn more about the distribution of certain venues. As you move further into the centre of London for example there are more stations and there begins to be London Underground stations which are much more densely situated. This distribution was not however used in the analysis due to other variables such as the commutability score which takes this into account and provides a deeper insight than this particular query.