**Applied Data Science Capstone Project - The Battle of Neighborhoods**

**Introduction/Business Problem**

As I am from London UK it seemed logical to focus this project on the London area. In this project I will attempt to develop a model to predict the median property value for a neighborhood in Inner London based on factors such as distance to central London, number of schools, doctors offices, parks, transport links such as train stations. The purpose of this is to understand the factors which influence property value in London so that the effect on property values of adding or removing infrastructure or amenities can be estimated and the cost/benefit of a change can influence decisions about public or private sector spending. It will also be interesting to see how changes in working practises following the Covid-19 pandemic may influence house prices in London.

**Data**

First, I will determine how to divide London into neighborhoods so that there are enough that statistically significant conclusions can be drawn but yet not so many that analysis is impossible. I will also need to use definitions of neighborhoods for which data is available concerning median property values.

There are approximately 322,000 detailed postcodes in London which can be rolled up into 276 ‘outcodes’. There are 32 Boroughs in London of which 12 are considered ‘Inner London’. There are 627 Wards in London of which 221 are in ‘Inner London’.

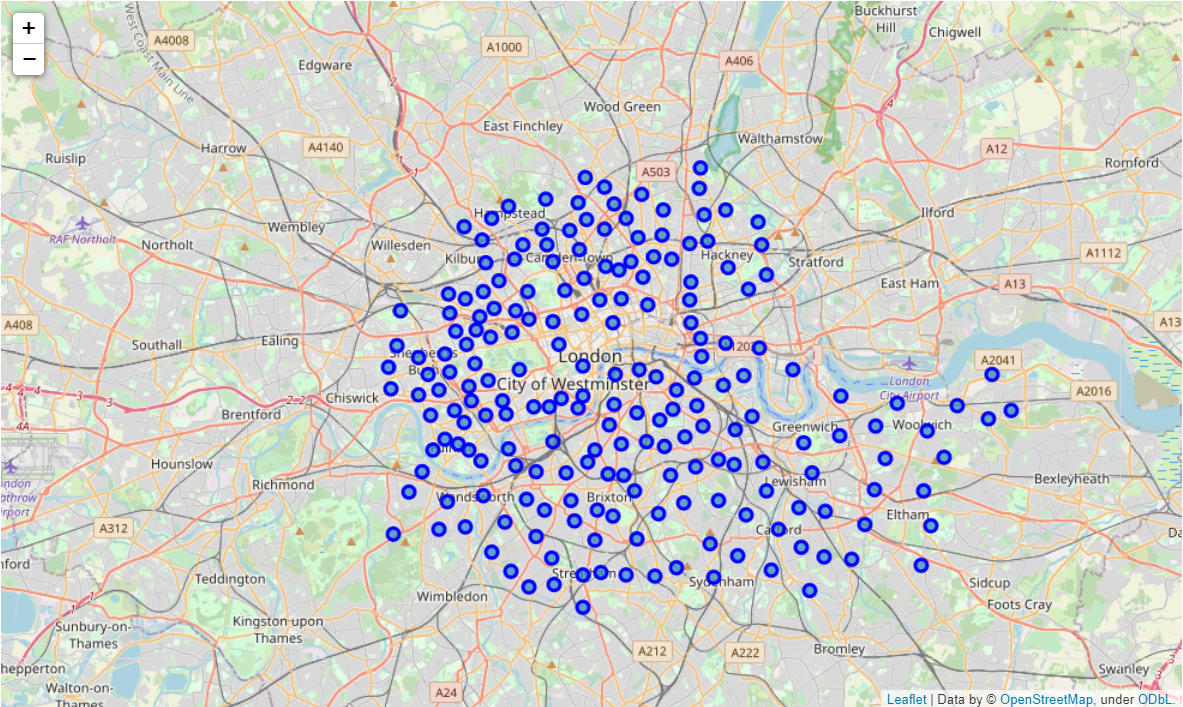
The ONS (Office of National Statistics) website is a good source for a dataset which includes longitude and latitude for electoral wards in the UK so I will use this to identify the geographical coordinates for the wards in London. I will take the median property value data from data.london.gov.uk which is published by the office of the Mayor of London. The distance to central London will be based on co-ordinate data and the number of schools, doctor’s offices and other amenities will be obtained from Foursquare.

Having decided that the best way to categorise neighborhoods in London was by electoral Ward, I next needed to create a dataframe comprising of a list of Wards in Inner London from which I could build the data table for my analysis.

I used the ward-profiles-and-atlas data from data.london.gov.uk and created a dataframe using only the wards that were associated with the boroughs identified as ‘Inner London’ in the ‘List of London Boroughs’ on Wikipedia. I then added the longitude and latitude of each ward to the dataframe by taking the coordinates from the wards-december-2017-full-clipped-boundaries-in-uk-wgs84 file on the Office of National Statistics website (<https://geoportal.statistics.gov.uk/>).

I had to remove a small number of wards as changes were made to ward names and boundaries in 2014 in the Boroughs of Hackney, Tower Hamlets and Kensington and Chelsea and the house price data predates these changes.

I visualised the wards on a map of London using Folium to makes sure that they are relatively evenly spread and covering a reasonable area.



I then added the distance of each set of coordinates from a point in Central London – for this project I chose Trafalgar Square as a proxy.

I looked for nearby venues using Foursquare and created a list of the top ten venue types for each ward to visually check that the wards appeared to have different characters and to check that my working so far gave sensible results.

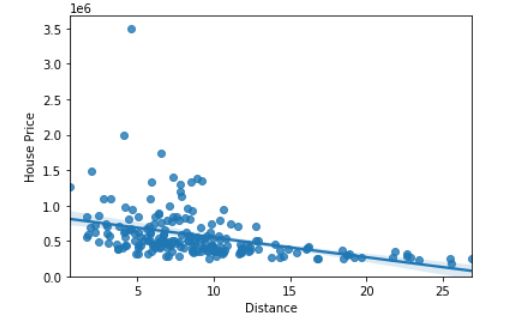
To generate the counts of specific types of venues, namely parks, doctor’s offices, schools and train stations I ran searches using Foursquare to create a list of venues for each category using category codes from <https://developer.foursquare.com/docs/build-with-foursquare/categories/>. I then summed up the number of venues returned for each ward and added them to my dataframe. If none were found I made sure to put a zero in the table.

Finally, I added the median house price for each ward to the dataframe, obtaining the data from the ward-profiles-and-atlas spreadsheet from data.london.gov.uk.

**Methodology**

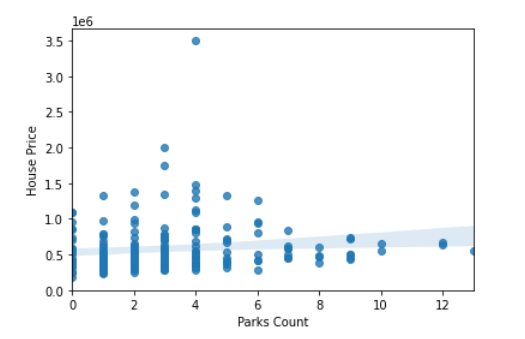
I first looked at the relationship between each potential predictor variable and the output, median house price.

First, I visualised the relationship between distance and house price using a Seaborn regression plot. The graph looked like this:



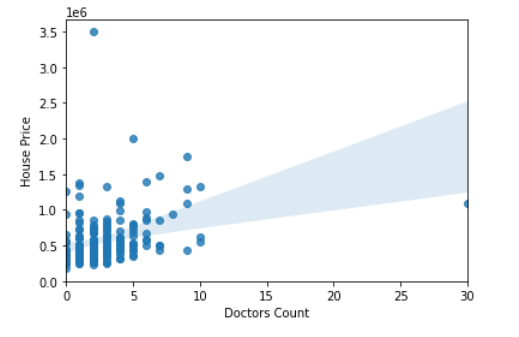
The Pearson correlation coefficient was -0.387 with a P value of 1.26 x 10-8

Next, I visualised the relationship between number of parks and house price using a Seaborn regression plot. The graph looked like this:



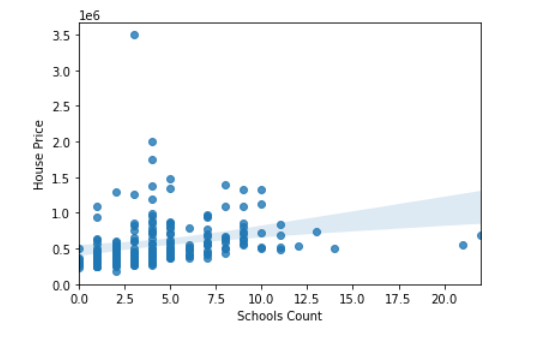
The Pearson correlation coefficient was 0.122 with a P value of 0.084

Similarly, the relationship between number of doctor’s offices and house price looked like this:



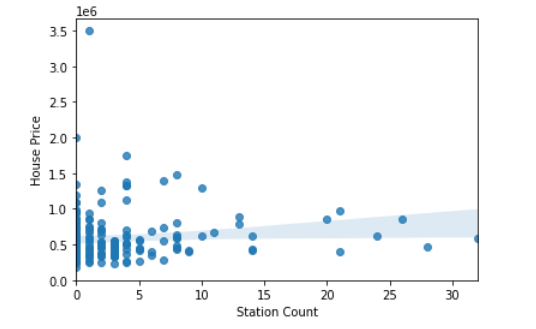
The Pearson correlation coefficient was 0.316 with a P value of 4.47 x 10-6

The relationship between number of schools and house price looked like this:



The Pearson correlation coefficient was 0.238 with a P value of 6 x 10-4

Finally, the relationship between number of train stations and house price looked like this:



The Pearson correlation coefficient was 0.097 with a P value of 0.169

**Results**

**Discussion**

Foursquare data is somewhat limited for a project such as this as it is crowd-sourced and therefore subject to inaccuracies as the accuracy depends on the data entered by users of the Foursquare app and the way that owners of businesses categorise their businesses.

**Conclusion**