**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’.

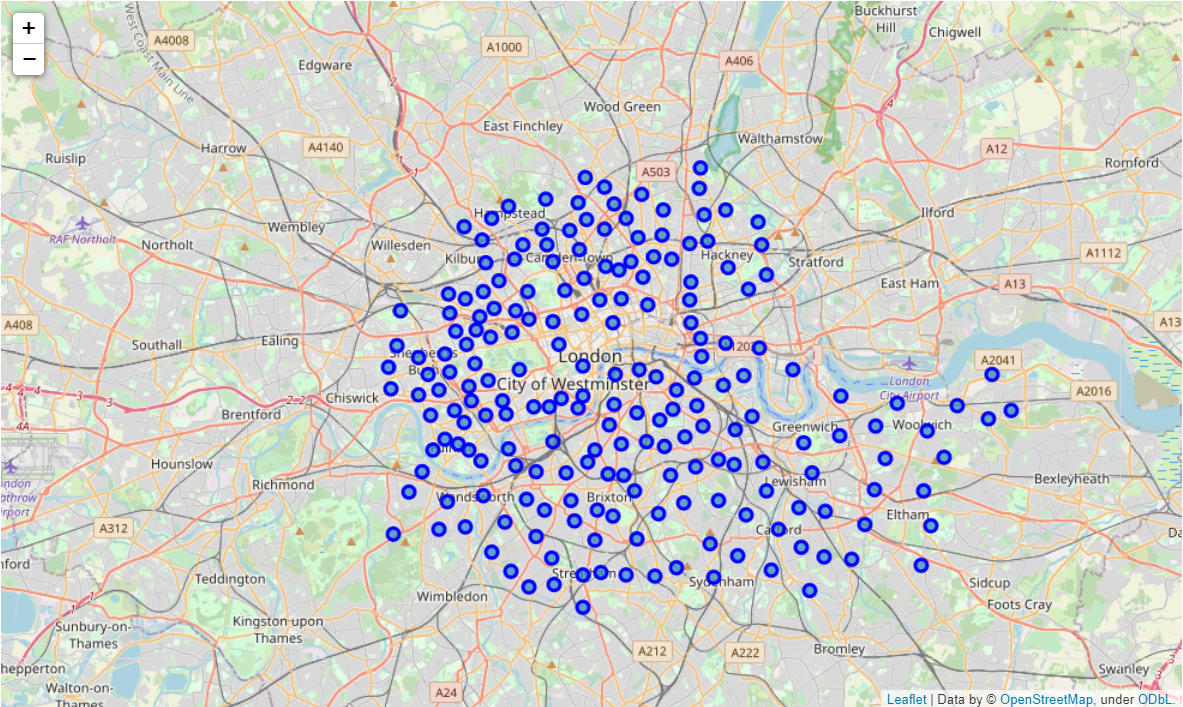
The ONS (Office of National Statistics) website is a good source for a dataset which includes longitude and latitude for electoral wards in the UK so I will use this to identify the geographical coordinates for the wards in London. I will take the median property value data from data.london.gov.uk which 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**

My plan was to look at the relationships between the predictor variables that I had selected – Distance to Central London, number of parks, number of doctor’s offices, number of schools and number of train stations - and the output that I had selected – median house price.

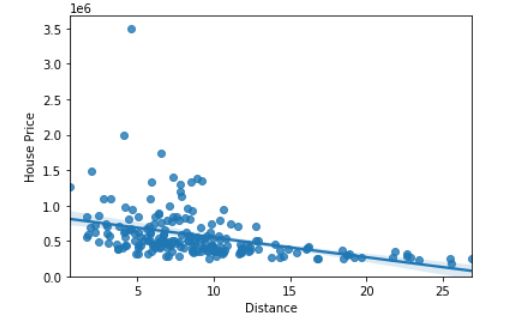
I first looked at the relationship using regression plots from the Seaborn library in Python.

I then constructed a linear regression model using the Scikitlearn library in Python for the Distance to Central London and a multiple linear regression model using the predictor variables that showed a statistically significant linear relationship.

Following this I constructed a polynomial regression model and trained it using a sample of the dataset then tested it using a smaller sample.

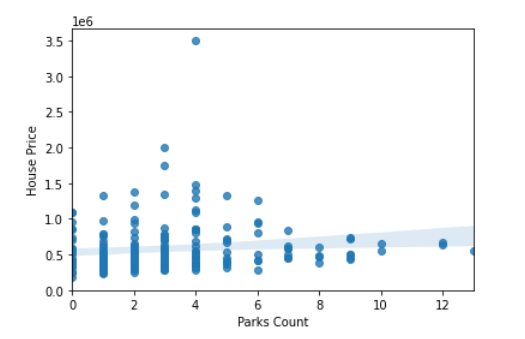
**Results**

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



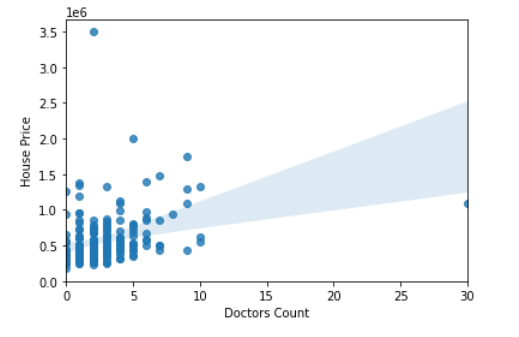
The Pearson correlation coefficient was -0.387 with a P value of 1.26 x 10-8

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



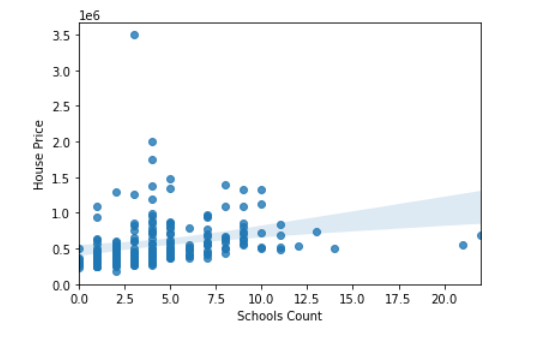
The Pearson correlation coefficient was 0.122 with a P value of 0.084

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



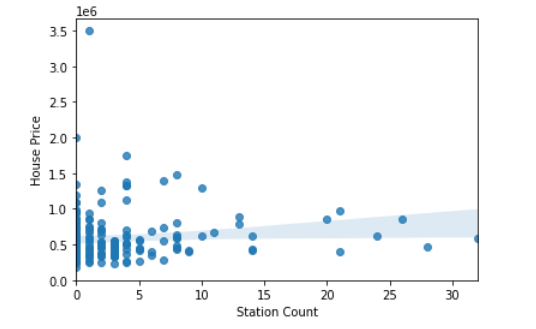
The Pearson correlation coefficient was 0.316 with a P value of 4.47 x 10-6

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:



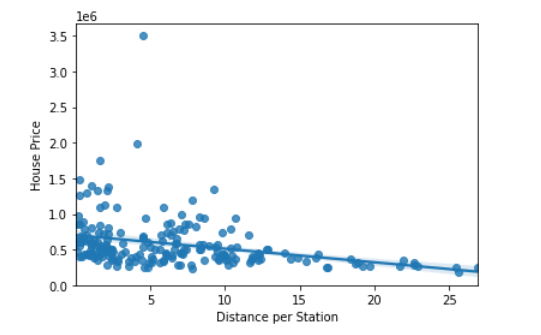
The Pearson correlation coefficient was 0.097 with a P value of 0.169

Distance to Central London appears to have a weak relationship to house price however it is statistically significant. The number of doctor's offices also has a weak relationship to house price which is statistically significant. The number of schools has a very weak relationship to house price which is statistically significant.

Neither the relationship between number of parks nor number of train stations to median house price were statistically significant as the p-values were greater than 0.05.

I suspect that a factor which influences house price is the time taken to travel to Central London which relates to distance and number of stations. In order to look at this I divided distance by the number of stations and then looked for correlation with house price to see if this improved the correlation. I was not sure how best to treat the data points where there are zero stations so I treated them as if there was one station.

The seaborn regression plot looked like this:



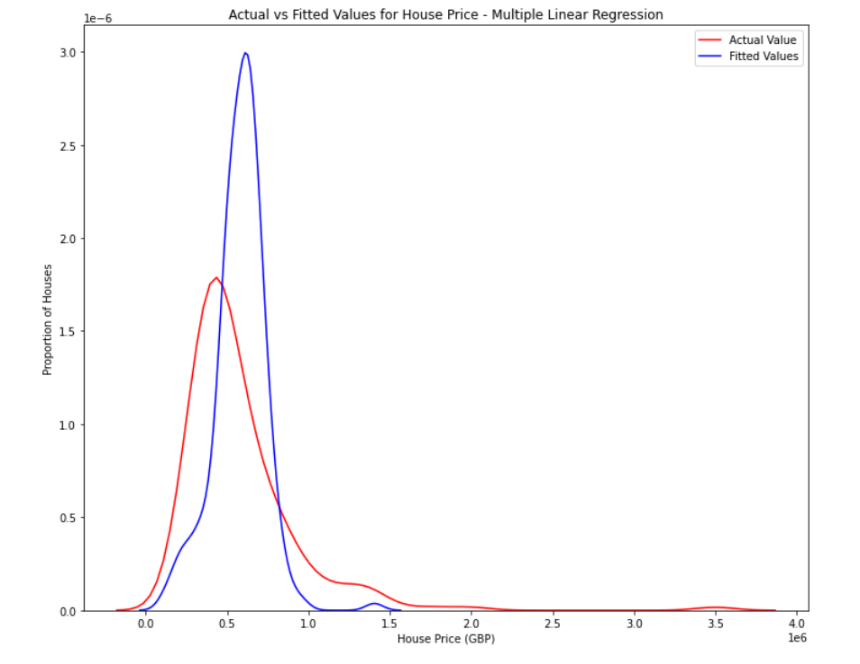
The Pearson correlation coefficient was -0.309 with a P value of 7.627 x 10-6

This was not obviously different to the relationship for distance to Central London alone so I concluded that that either bringing the number of stations from Foursquare into the calculation would not make the factor more representative of the time to travel to Central London or that the ease of travel to Central London is not a significant factor affecting house price in London. I therefore chose to continue the data analysis using distance as a factor.

I then set out create models to predict house price using the variables that I had identified were statistically significant individually. I first created a Linear Regression model using the Scikitlearn library in Python to examine the relationship between distance to Central London and house price.

The R squared value for this model was 0.150 suggesting that distance to Central London explained only around 15% of the variation in house price.

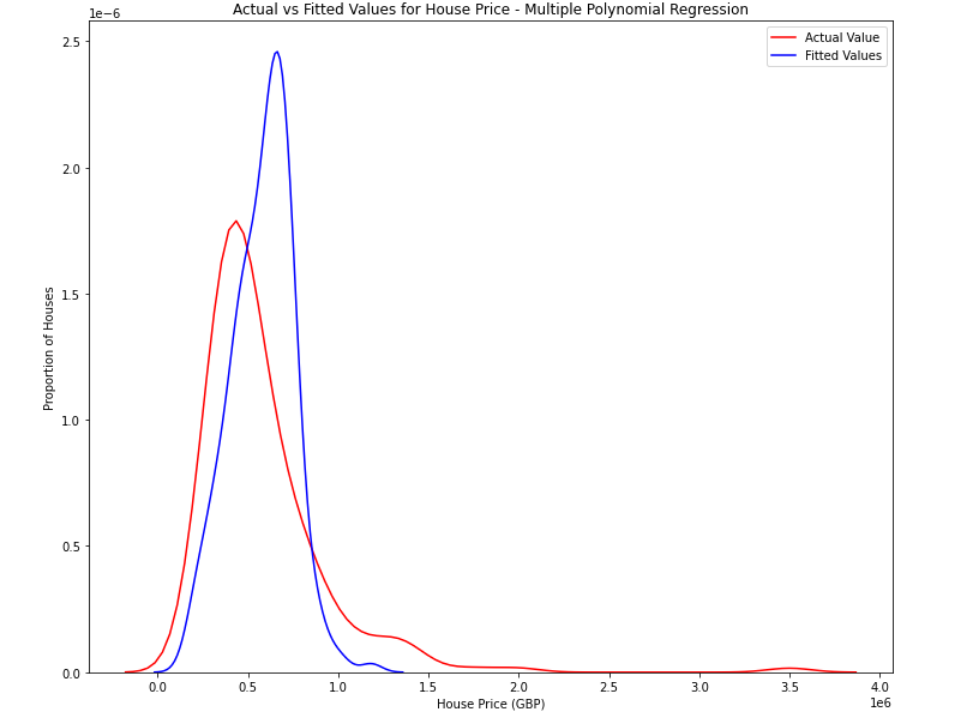
I next created a multiple linear regression model using Distance, Number of Doctor’s Offices and Number of Schools. The R squared value for this was 0.191 suggesting that these factors explained around 19% of the variation observed in house price. I plotted the actual values versus predicted values to see how closely the model predicted observed values and obtained the following chart:



The model is not a good predictor of the observed values as we might expect from the low R squared value.

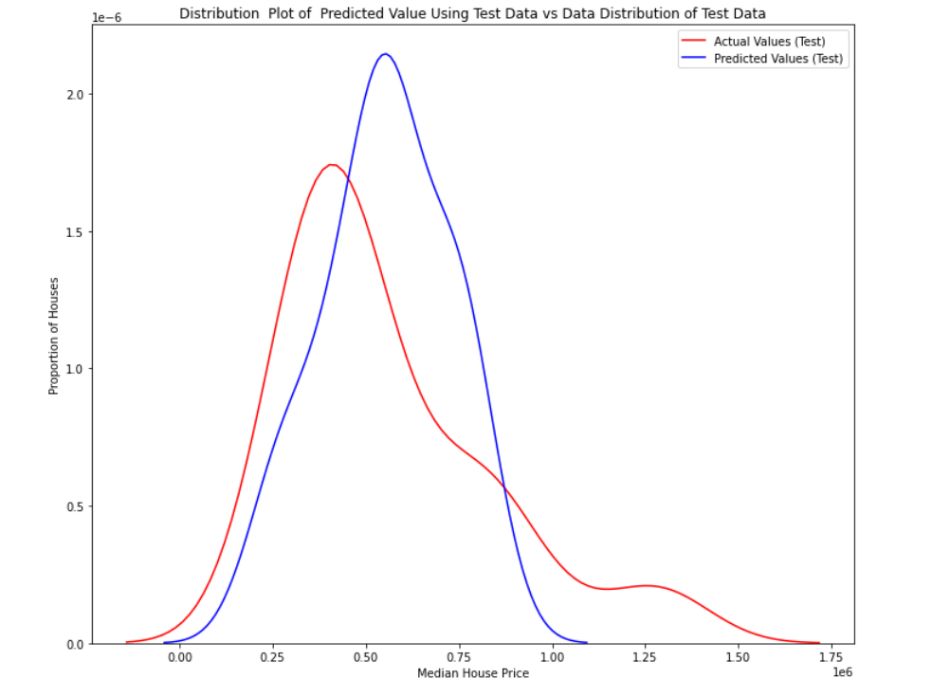
In order to investigate whether a polynomial model would give better results I constructed a multiple polynomial model for the predictor variables in the second degree using Scikitlearn’s PolynomialFeatures function. I first carried out a polynomial transformation and fit in the second degree.

The multiple polynomial model produced an R squared value of 0.228 which is an improvement on the multiple linear regression model. The plot of actual values versus predicted values was as follows:



As a final exercise I split the data into a training set and a test set and re-run the analysis to verify the model. I used 80% of the data for training and 20% to test the model.

The R squared value for the training data was 0.215 while the R squared value for the test data was 0.296. The plot of distributions for the test data was as follows:



**Discussion**

The relationships between the predictor variables and the output do not appear obviously linear so it is no surprise that the polynomial model gives a higher R squared score.

The polynomial model is a poor predictor of house price so further work would be necessary to identify additional predictor variables and to refine the model. I suspect that relationships between the predictor variables and the output are far more complex than this fairly superficial study could do justice.

The London property market has many nuances – distance to Central London has traditionally been a factor however when you move further away from Central London the commuting time is more of a factor so proximity to stations with fast connections to Central London will be more desirable and therefore have higher property prices. Property prices have risen locally in areas which will be on the new Elizabeth Line rail network (also known as Crossrail) which will connect suburbs to Central London with a fast commuter rail service. How this can be modelled would need to be studied further.

There are limitations to what can be learned from Foursquare data 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. The categories are in some cases too granular and in other cases too high level to allow meaningful conclusions to be drawn. For example, proximity to certain schools is also a potential factor however not all schools are created equally and Foursquare cannot easily be used to distinguish between schools.

The Foursquare data used in this study was obtained using the API during August 2020. As the data is crowd sourced it is constantly updated and so will change from time to time. The house price data is from 2014 and so may be somewhat out of data as since then London property prices have risen significantly and then fallen back.

**Conclusion**

Distance to Central London, Number of Doctor’s Offices and Number of Parks have a relationship to the median house price for a Ward in Inner London however the relationship is complex and will require further study to identify the individual relationships and interactions and additional factors will need to be identified.

It is also apparent that Foursquare data has limitations for obtaining data to carry out a study of this type.