What are the determinants of London house prices?

Introduction

The factors that influence house prices in London are of interest to both individual investors looking to purchase property to live in or as an investment as well as private companies who intend invest in property for purely financial gain. If the effect of on more of these factors could be identified it could be used by both groups to improve the quality of their decision making in this process. This project will investigate the effect of 4 factors on average house prices, these will be path dependence, population by borough, average life satisfaction by borough and number of schools by borough.

Data

The data used will come from three sources, the first is the website for HM Land registry which is a non-ministerial department tasked with registering ownership of land and property in England and Wales. This will be used for the data for dependent variable, average house prices by borough and path dependence for which the average house prices by borough from the year before will be used. The second will be London Datastore which is an open data sharing portal run by the Greater London Authority. This will be used for the data relating to population, life satisfaction and number of schools. The third data source will be Google maps which will be used to obtain the location data using its nearby function.

As both London Datastore and HM Land Registry are run by public sector organisations the data is both accurate and reliable. Also, the location data from Google maps is reliable as it is an established and market leading company in its sector. Due to data availability issues all data will be for the year 2015 except for the data for average house price by borough 2014 and the location data.

Data Source	Data Set	Variable
HM Land Registry	average house prices - borough	Average house price by borough 2015
HM Land Registry	average house prices - borough	Average house price by borough 2014
London Datastore/ONS	Office for National Statistics (ONS) Population Estimates, Borough and Ward	Population by borough 2015
London Datastore	personal well-being - Borough	Life satisfaction by borough 2015
London Datastore	schools pupils borough	Number of schools by borough 2015
Google Maps	Individual locations data using the nearby function	Latitude and Longitude

Methodology

This project will use a quantitative approach when analysing the factors that affect average house price by borough. This is because the relationship between the independent variables and the dependent variable is quantifiable. As all the data to be used in the regression except for the data for path dependency is from 2015, a cross section analysis will be used. This is because a cross section analysis allows the effects of dependent variables to be measured at a specific point in time. The cross-section analysis will be carried out using multiple linear regression as there are multiple independent variables. The independent variables except for average life satisfaction by borough have been chosen as they are all commonly believed to have some effect on the average house prices of a region. The variable average life satisfaction by borough is being included to see if it captures the effect of multiple factors such as the population, total air pollution and crime rate that could all also impact the average house price of a borough. This variable will be measured using an average score for each borough with a range from 0 to 10.

To complete this project the individual data sets will be combined in Microsoft excel using a VLOOKUP formula with the borough name as the unique identifier, this will then be saved as a csv file so that it can be imported in a Jupyter Notebook. Then the data types will be checked using python and will be updated if necessary, so that they behave as required in the following data visualisation and analysis. After this, the different borough will be visualised using folium to show the different locations of the boroughs within London. The data will then be visualised again using Scikit-learn to product a scatter plot for each of the independent variables with a line of best fit to see if there are any initial observations that can be made. Finally, multiple linear regression will be used to calculate the coefficient of each of the independent variables to determine their effect on the dependent variable.

Results - Data Visualisation
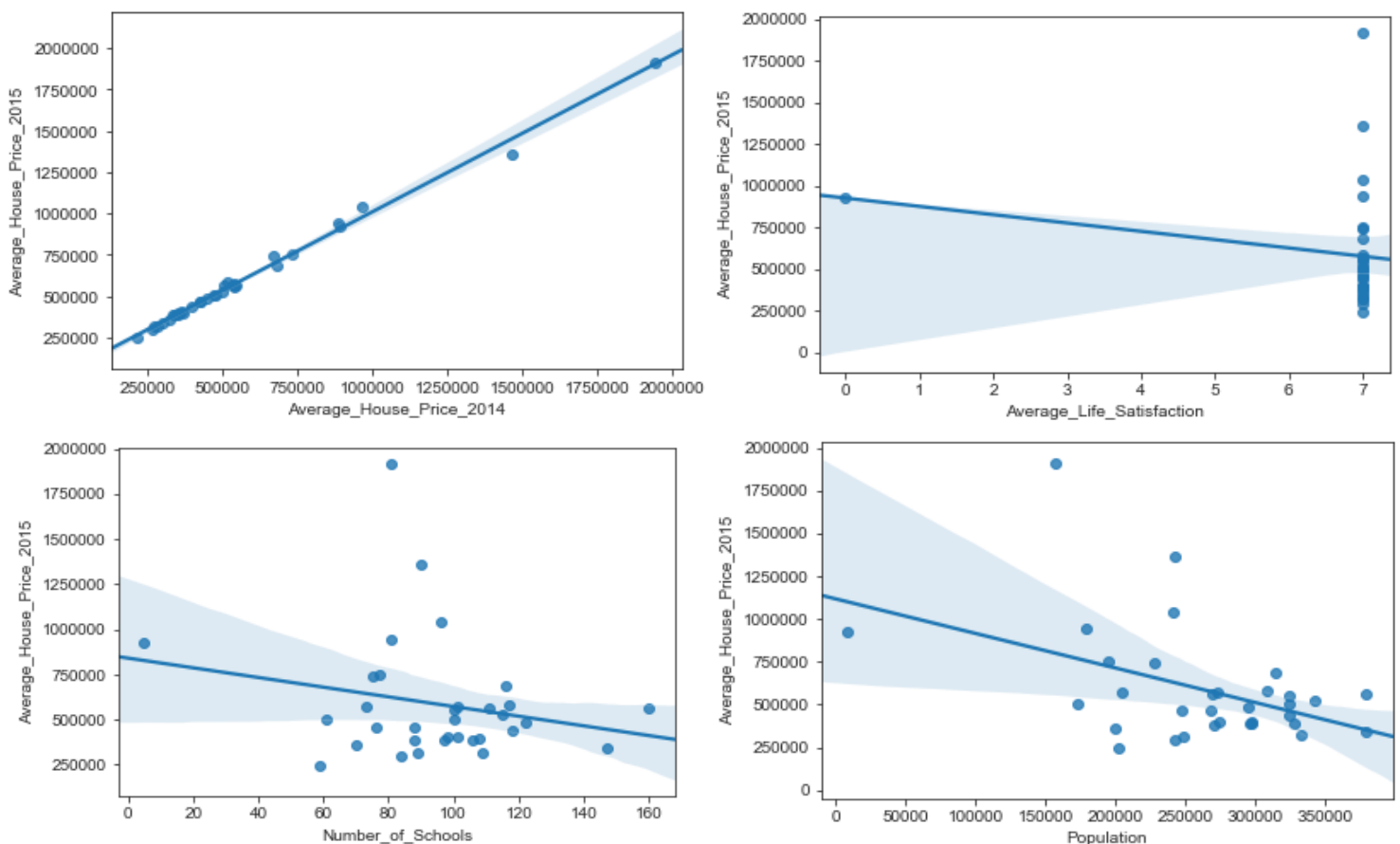
The first visualisation that has been completed suggests that there is clearly a positive relationship between the variables for average house price for 2014 and 2015. The second visualisation suggests a negative relationship between average house price and average life satisfaction. The third visualisation suggests that there is negative relationship between the number of schools in a borough and average house price. The fourth visualisation suggests that there is a negative relationship between the population of a borough and average house price.

Results - Multiple Linear Regression

Independent Variable	Coefficient
Average house price by borough 2014 – path dependency	9.28538281e-01
Average life satisfaction by borough	-8.48651061e+02
Population by borough	-3.50540899e-01
Number of Schools by borough	7.94095987e+02

The results of the multiple linear regression show that the coefficients for path dependency and number of schools is positive and that the coefficients for average life satisfaction and population are negative.

Analysis of Results

Path dependency as shown in both the regression and the scatter plot has a strong positive relationship with average house price. Average life satisfaction by borough is shown in both the scatter plot and the regression to have a negative relationship with average house price. It should be noted that the scatterplot displays an outlier where the average satisfaction is 0. This observation is for the City of London where no data was collected. A potential way to improve the accuracy of the regression would be to remove this observation and repeat the regression. However, as almost all the data points for this variable are 7 out of 10 the coefficient of this variable will be disregarded as it does not have enough range for its effect to be identified accurately.

The regression coefficient for the number of schools is positive meaning that there is a positive relationship between the number of schools and average house price. Despite this the scatter plot for this variable has a negative gradient suggesting a negative relationship between the two variables. A possible explanation for this could be that including multiple variables in the regression changed the effect of this variable as including more independent variables in statistical tests improves accuracy.

Population is shown to have a negative relationship with house prices in both the regression and the scatter plot. When both the scatter plots for population and number of schools are compared it seems as though the effect of both variables is similar. A possible reason for this could be that there is collinearity between these two variables. For example, the population of a borough could influence the number of schools in a borough. However, as the coefficients from the regression completed after the scatter plots were produced shows that population and number of schools have opposite effects, I do not deem this to be an issue.

<u>Conclusion</u>

This project shows that path dependency and number of schools by borough have a positive relationship with the average house price by borough. This means that when an individual or a company are deciding which borough of London to buy property in. They should consider that the higher the number of schools and the higher the average property prices in previous years for a borough the higher the average price of a house will be.

The effect of average life satisfaction by borough was inconclusive due to a lack of range within the dataset. Population by borough is shown to have a negative relationship with the average house price by borough. This means that individuals and companies should expect borough of London with higher populations to have lower house prices.