NCG 613 DATA ANALYTICS PROJECT

Lecturer

Professor Kevin Credit

Group Project on

LONDON BOROUGHS HOUSE PRICE ANALYSIS

Student Name	Student Number
Jayasurya Duraisamy	21250461
Naresh Kumar Nagaraj	21250702
Sanley Pathrose Oommen	21250765
Sreevathsa Devanahallibokksam	21250214

Date: 13.05.2022

CONTENTS **Page No.**

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This study examines the patterns in the prices of the London Borough area with the help of geo-spatial data, hedonic modelling, and supervised machine learning tools. The study represents k-nearest neighbour regression technique to predict the house prices based on few significant predictors such as geometry, Dist_school, floor area, type of house etc. The accuracy of the model was calculated to 67.15%. Hypothesis test have been performed to check whether the house price is significantly affected by the distance to nearest School.

1. INTRODUCTION

The 32 local government areas that make up the ceremonial county of Greater London are known as London boroughs, and each is managed by a London borough council. The current London boroughs are a form of local government area that were founded at the same time as Greater London on 1 April 1965 by the London Government Act 1963. Out of the 32 boroughs, twelve are designated as Inner London boroughs and twenty as Outer London boroughs. The historic centre of London is a separate ceremonial county and unique local government unit that operates differently from a London borough. However, the two counties together make up the Greater London administrative region, as well as the London Region, which are both overseen by the Greater London Authority. The historic centre of London is a separate ceremonial county and unique local government unit that operates differently from a London borough. However, the two counties together make up the Greater London administrative region, as well as the London Region, which are both overseen by the Greater London Authority.

Inner London boroughs tend to be smaller, in both population and area, and more densely populated than Outer London boroughs. The London boroughs were created by combining groups of former local government units. A review undertaken between 1987 and 1992 led to a few relatively small alterations in borough boundaries.

The impact of the housing crisis on most Londoners is unavoidable. Too many families have been unable to find a genuinely affordable place to call home in London due to a toxic mix of rising private rents, low housing benefit levels, and the elimination of social rent housing. Many Londoners are now unable to pay their rent and are forced to live in overcrowded or inappropriate circumstances.[1] (Figure 1) shows the total house constructed vs year. We can observe that, the construction of houses has kept on decreasing from the year 2011 to 2019. Home ownership is a faraway goal for many in this age. To overcome this, government came up with an idea “London Housing Strategy” which was adopted in August 2018. The aim of this strategy is to address the housing shortage through an intensive use of London’s available land, focusing on more genuinely affordable housing[2]

1 More info on housing crisis at https://blog.shelter.org.uk/2020/02/a-capital-in-crisis-what-has-caused-londons-housing-emergency/
2 More info about LHS at https://www.london.gov.uk/what-we-do/housing-and-land/tackling-londons-housing-crisis/

Apart from the internal factors of house such as total floor area, number of rooms etc. There are various external factors which influence the house price such as community, transportation, access to schools, universities, hospitals etc.

School accessibility is frequently acknowledged as an essential factor in determining house pricing. It is frequently mentioned as a significant selling element for new residential buildings. Data from the National Association of Realtors shows that 26% of recent homebuyers are influenced by school district when selecting where to live[3]. That percentage jumps to 46% for buyers ages 30 to 39, and by 36% for buyers ages 22 to 29. On average, a one percentage point increase in the neighbourhood proportion of children reaching the government-specified target grade pushes up neighbourhood property prices by 0.67%.

Most international purchasers are unconcerned about the potential impact of Brexit on the UK as a destination to educate their children, and there are even signs that it is becoming more appealing to some[4].

The analysis looks the trends in the prices of house from the year 2018 in the London boroughs area influenced by distance to school.

2. LITERATURE REVIEW

The UK House Price Index (UK HPI) measures changes in residential property values. The UK HPI is based on sales data from residential housing transactions in the United Kingdom, whether for cash or with a mortgage. Properties have been included in England and Wales since January 1995. Data is available at a national and regional level, as well as counties, local authorities and London boroughs. The UK HPI has a substantial data source (land registrations such as those kept by HM Land Registry) that allows data to be released down to a local authority level, with further breakdowns available by property type, buyer status, funding status, and property status. (GOV.UK, n.d.)

However, the limitations of House Price index are it is not as timely in publishing as other house price index measures published in the UK because it is based on completed sales at the end of the conveyancing process, rather than advertised or approved prices

3 More Info about how property values affected by schools at https://www.upnest.com/1/post/property-value-affected-by-school-district/
4 More info about overseas buyers at https://www.theguardian.com/uk-news/2018/sep/05/wealthy-overseas-parents-london-property-private-school-places

The relationship between schools and house prices has been studied extensively. Housing prices were regressed on distance to schools while controlling for house and community factors in previous works.

Black's (1999) research, which is one of the most widely cited, looked at how property values in Massachusetts changed when they were in the catchment area of an excellent school. In neighbourhoods separated by this line, Black looked at property values within 0.15 miles of the attendance boundary of the catchment regions, because children from the same neighbourhood will attend different schools. Thus, neighbourhood features are taken into account. After controlling for the neighbourhood impact, the price disparities between comparable residences on different sides of the boundary would be due to school quality. (Black, 1999)

The concern with this concept is that, owing to potential future changes in the placement of the boundary, housing values near the boundary may be lower than house prices closer to the school. (Yinger, 2011)

Various researchers have studied about how the quality of school (performance) which is measured by various factors such as how the students have performed in Reading, writing and maths test and have calculated the average of performance of school. Based on this they have rated the quality of the respective school.

Stephen Gibbons' (2012) research, which drew on Black's (1999) paper, showed comparable results by matching identical features across authority borders using an enhanced boundary discontinuity regression model. They discovered that a change of one standard deviation in school average value-added or prior accomplishment raises the price by about 3%. (Stephen Gibbons, 2012)

All of this research has considered only the middle school and secondary school and have tested the quality of the schools. Our research is based on considering all the schools in overall (primary, middle and secondary) school as parents with children of different age category would like to send their children to the nearest school from their house for ease of travel and in concern with safety of their child. So, we have selected London all school’s dataset which consists of all the schools in the London borough

In this project, we will utilize the Hedonic Regression model to examine the behaviour of housing prices in this study. Hedonic modelling is a technique used in geographic data analysis to analyse the impact of multiple factors that might contribute to house pricing.

3. DATASET ANALYSIS

Following are the datasets used and analysed to predict the house prices.

- London-borough-house-price data

  The dataset is comprised of 6411898 rows and 17 columns. The dataset has house price information of houses specific to London boroughs. Most significant columns are price and geometry. Few of the columns are independent, whereas others are dependent. For instance, 'priceper' is obtained from the columns price and tfarea (total floor area). Year column represents the date of sale as mentioned in sale of deed. The detailed feature description of this dataset can be found on (Table 1). The number of houses in Inner boroughs are more than the number of houses in outer boroughs.
- London-boroughs data

  The dataset is comprised of 33 rows and 7 columns. The dataset consists of information regarding the districts constituting the London boroughs area such as boroughs name, Hectares, and geometry. The geometry column is a Multi Polygon data. This could be used to get the location of houses by boroughs name or with the boundaries. Refer to the (Table 2)

- School dataset

  The dataset consists of 3242 rows and 7 columns. It has information about the schools such as the geometry (which is a significant column for our analysis), borough name where it is located, Post code etc. (Table 3) shows the description of features.

  Here, we have calculated the nearest distance to school from the house. Now, this variable is the one on which we have conducted hypothesis test to check whether it is a significant factor which influences the house prices. So, we have added this variable as Dist_school to our London-borough-house-price data. From (Figure 6), we can observe, that Kensington and Chelsea, Hackney have high density of schools.

- OutputAreas (with tree density, student density, and professionals' density) and a dataset on the distance of house from transit, center location, and distance from road near the houses were also employed. These data sets form hedonic relationship to the main data set of London boroughs.

4. DATA PRE-PROCESSING

In London borough house price dataset(houses_lr.geojson), there are some categorical variables (property type, oldnew, duration, categorytype, recordstatus, pcd and pcd2). Out of these features, we estimate that property type, oldnew and duration are important predictors which influence the house price. So, these values have been mapped to dummy variables to 0 and 1.

As a next step, we checked the data distribution of the features. While checking the distribution of our target variable ‘price’ and total floor area ‘tfarea’ it can been seen that both are skewed rightly as can be seen on (Figure 7) which will not provide a better fit, so to avoid this, it has been transformed by taking natural log of the variable. Violin plot of the transformed variable can be seen on (Figure 8) which is much more conductive to a linear fit with other normally-distributed variables.

With the help of openspace.geojson, houses_roads_2018.geojson and TransitStations.geojson the nearest openspace, road and transit station has been identified. Similarly, for the school’s dataset, for every house we have calculated the nearest school and filtered the status of the schools which are currently operating. And these distance features are included in the dataset by spatial join. Because our housing year should be after the construction or establishment of covariate (in our case it is schools), we filtered out the house price dataset to 2018.

5. HYPOTHESIS TESTING

Hypothesis test was performed on the data to check the significance of predictor variable.

Null Hypothesis(H_0): Distance from house to nearest school has no significant effect on the house price.

Alternative Hypothesis (H_A): Distance from house to nearest has significant effect on the house price.

The hypothesis data includes a variety of features, but the year has been confined from 2018 as the housing year should be after the year of establishment of our covariate. The predictor variables used for testing the hedonic model are the price of the house ('price') , area of the house ('tf_area'), 'number of rooms', the type of house('flat', 'detached'), duration of the house ('leasehold', 'Freehold'), distance to nearest school ('Dist_school'), distance to road ('Dist_Road'), distance to transit ('Dist_Transit'), Distance to the centre location('Dist_0KM') and percentage of households with at least one deprivation dimension ('DEPRHH'). To test the hypothesis various models were fitted, and the values were compared reaching to the conclusion.

6. MODELS USED FOR HYPOTHESIS TESTING:

Ordinary Least squares: OLS is used in estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable. Such kind of models are not suitable for analysing the spatial data as there can exist spatial dependence and spatial errors driving the model resulting in inaccurate predictions. Also, the assumptions of linear regression such as Normality, Linearity, collinearity, Homoscedasticity fails in most of the spatial data. General expression of linear regression can be given by:

$$Y = X\beta + \varepsilon$$

Where Y is the natural log of price and x is the vector of housing features such as logarithmic value of floor area, number of rooms, property type, duration, distance to school etc. e is the error term which tells us the difference between the actual value and predicted value (in our case it is price) and b is the vector of coefficients of predictors. Summary table of OLS can be seen in Table 5(Table 5).It can be observed from the table that all the features are significant and

the features which are related to distance has negative effect, which means that as closer the house, the price will be higher. 'Dist_school' is a significant prediction has a negative effect if all the other predictors are kept constant.

Detecting if geographical variables are playing a hidden role in causing variance in home pricing is a crucial component of knowing how to estimate house prices. This may be accomplished by spatially combining the home pricing data set with the Output Area data set and then showing the residual distribution by boroughs. It can be observed from the (Figure 9), a significant variation in the residual prices can be observed per borough. The observed values for home prices in the boroughs on the left of the scale are much lower than the model would anticipate based on the actual housing attributes submitted to the model. Similarly, the prices in the boroughs on the right are higher than the model predicts for the given set of housing attributes. This might be related to the fact that most of the boroughs on the left of the scale are spatially close to one another and hence cluster. This can be tested by creating a distance-based spatial weights matrix and generating spatial lags of the residuals and then correlating these spatially-lagged residuals with the residuals of the model in their origin locations. From Moran's I plot of model residual (Figure 10), It can be clearly seen that there is a strong positive correlation between the residuals of the model in the given plot and the residuals of the model in its neighbours. It help us infer that there are spatial processes been plotted(Figure 11). Here, red points correspond to "high-high" spatial clusters of positive residuals, while blue points correspond to "low-low" spatial clusters of negative residuals. The red points are mostly being plotted at the center of London while the blue point occupies the other regions.

Spatial Lag Model: It is used in cases where there is spatial dependence on dependent variable (i.e.) events in one place predict an increased likelihood of similar events in neighbouring places

$$Y = WY\rho + x\beta + \varepsilon$$

Where W is the spatial weights matrix, ρ is the spatial lag coefficient. This can be interpreted as prices of houses which were set based on house prices of the neighbouring houses.

From summary table (Table 6), it can be observed that the spatial lag coefficient is positive and significant (probability 0.000), indicating positive spatial dependence in the dependent

variable. However, based on the spatial diagnostics above, we might still assume that there is some residual spatial error in the model.

Spatial Error Model: Spatial Error model assume that error terms in regression are correlated. This method is used when there is some spatial residual error. The General equation of this equation is:

$$Y = X\beta + \lambda W\upsilon + \varepsilon$$

we split this error into random error and spatially structured error. Here, random error (unexplained by our model) is u, independent identically distributed (i.i.d.) errors. Our spatially structured error is composed of the spatial error coefficient λ, and our original ϵ error term now weighted by our weight matrix w. From Summary table of this model (Table 7), it can be observed that spatial error coefficient λ is positive and significant (Probability: 0.000) which would have positive impacts on other variables in our analysis.

Spatial Lag + Error Model: The Spatial Lag and Spatial Error models are combined in this model. It includes elements from both the Lag and Error models**.** (developers, 2019)

$$Y = \rho WY + X\beta + \lambda W\rho + \varepsilon$$

It is commonly assumed that just one form of spatial dependence is important, and that the spatial lag + error model should be used to test the impacts of the other models. From summary of the model (Table 8), it can be observed that 'W_log_price' is no longer significant and we can interpret that 'lambda' is positive and significant (Probability=0.000). Hence, this suggests that spatial error would be the better model to explain the variation of prices.

As we select spatial error model, hypothesis testing was performed on this model. We can see that the p-value of 'Dist_school' is 0.0000316 , which states that the feature is more significant for our model. So, we reject the null hypothesis at both 95% and 99% Confidence Interval

7. PREDICTION MODEL FOR THE HOUSE PRICES

A model was built to see if it was possible to ploict the price of a house in the London boroughs area. To predict the house price, the K-nearest neighbor model was chosen. Because the data set had a large number of attributes (17), the unnecessary independent features were removed from the dataset which includes "postcode," "dateoftransfer," "recordstatus," "priceper," "pcd," "pcd2," "dointr," "doterm," "GSS CODE," "geometry," "Detached," "Flats," and "New," which could be obtained from other attributes or did not play a significant role in the prediction process. To change the data in accordance with the model, data manipulation was required. Because the KNN regressor model does not account for categorical predictor variables, 'propertytype', 'oldnew', 'duration', 'categorytype', are all important features to the prediction model, but they are categorical. So, the data was mapped to dummy variables to 0 and 1. Also, the geometry column x and y was reduced to a single column ‘geometry’ with x as latitude and y as longitude. For easier interpretation, the house's price was also divided by 1000. The accuracy of the model was observed to be 79.88 percent, which is better, when the data set was split into an 80 percent training and 20 percent test set with 20 neighbours, distance as a weight parameter, and considering the Euclidean distance between the points. 10 KNN models were executed with different sets of parameters and cross-validation sets to see how the model generalizes. Cross-validation is a model validation technique for determining how well the results of a statistical analysis will generalize to a different set of data. The model with highest accuracy is 67.15% (Table 4).

8. RESULT

The data set for London borough house prices and the data set for London schools were used to conduct the experiment and test the hypothesis. Shape files for London were also used to plot the boundaries between the various boroughs in context. To understand the spatial density of the plot, all available houses from 2011 to 2019 were plotted on the map using the data set. (Figure 2) shows a scatter plot of all the houses on the map, coloured according to the logarithmic value of the price variable. This is done to normalize the price before graphing it. The City of London, as well as central boroughs such as the City of Westminster and

Kensington and Chelsea, have higher prices (highlighted in yellow and light green) than the other areas under consideration.

In comparison to the outer districts, the inner districts are smaller and have a higher population density. In the London Boroughs, (Figure 3) depicts a clear distinction between the inner and outer districts. From (Figure 4) we can interpret that the data is rightly skewed, mean is shifted towards the left of the median. This tells us that maybe lower house density observations have higher proportions. As can be seen by the dark blue highlights, both boroughs (Kensington and Chelsea, Islington) have a high density of houses.

From (Figure 6), we can observe, that Kensington and Chelsea, Hackney have high density of schools. This tells us that maybe lower house density observations have higher proportions and that this can strengthen our analysis that schools do play an important factor in houses construction, which could also lead to significance in their prices.

A comparative study of the mean and median house prices among the London Boroughs were conducted which can be visualized in (Figure 5). The average of all the houses in a borough was gathered and plotted on a map. This was done for each of the boroughs until a final heatmap of the boroughs, as shown in (Figure 5) (a), was obtained. The mean, on the other hand, is not always the best way to analyse averages because it is influenced by the highest and lowest values in house prices. (Figure 5) (b)shows the median of house prices plotted on a heatmap as an alternative.

The predictor dist_school is significant and has a positive effect on the house price, according to hypothesis testing on the spatial Error Model (Table 7). The predictive analysis KNN regression model correctly predicts the house price with a 67.15 percent accuracy.

9. CONCLUSION

Spatial house planning is a difficult regulatory process. When considering environmental and social considerations, determining realistic pricing can be difficult. In order to effectively estimate home prices for a particular location, several spatial features for a specific property needs to be included during modelling for house price estimates.

The dataset's exploratory data analysis found that median property prices were higher in the city center and significantly lower in the boroughs further out. Kensington and Chelsea

had the highest median property prices in London. One of the key causes of these swings was the rules surrounding Brexit.

In this project, a hedonic model was created to explore how the various variables in the dataset are spatially related to one another. It helped us figure out which variables in the dataset are influenced by their surroundings. For hypothesis testing, the null hypothesis was stated as the dist_school feature has no significant effect on house price and the alternative hypothesis as it has significant effect on the house price. With further analysis, it was found that the spatial error model was more appropriate in representing the dataset's hidden spatial dependencies. By further analysis with implementing spatial lag + spatial error model, it was observed that the dist_school variable was significant and had a positive impact on house prices. Features such as dist_tranit and dist_open becomes insignificant in almost all models as the probability value is more than 0.05 at 95% confidence interval. This may be due to the influence of size of the houses, as these types of houses may not be dependent on these features.

Even though the distance to school was a significant factor in determining the prices of houses across all boroughs, there may still be other factors that influence house pricing. This could be a research topic, and it could be combined with other existing resources to see if house prices have an impact on other environmental or socioeconomic factors. Based on the data analysis, it can be concluded that because the distance to school variable has a positive impact on house prices, government policies can be reformed to ensure that houses are priced fairly.

TABLES AND FIGURES

S.No	Feature Name	Description
1	Postcode	Postcode of the area where the house is situated
2	Price	Sale Price on transfer deed
3	Dateoftransfer	Date when sale was completed as stated on transfer feed
4	Propertytype	D = Detached, S = Semi-Detached, T = Terraced, F= Flats, O=others
5	Oldnew	Refers to the age of the property and applies to all price paid transactions, residential and non-residential
6	Duration	Refers to the tenure: F= Freehold, L = Leasehold etc.
7	Categorytype	Refers to the type of Price Paid transaction. A=Standardard Price Paid entry, includes single residential property sold for value B = Additional Price Paid entry including transfers under a power of sales/repossessions, buy-to-lets and transfers to non-private induviduals
8	Recordstatus	Indicates addictions, changes and deletions to the records. A = Addition C = Change D = delete Note that, where a transaction changes category type due to misallocation (as above) it will be deleted from the original category type and added to the correct category with a new transaction unique identifier.
9	Year	The year of house when it was sold as stated on the sale deed
10	Tfarea	The total floor area of the house
11	Numberrooms	Total Number of rooms in the house
12	Priceper	The per sq-meter price of the house
13	Pcd	7-charcter type of postcode (eg: ''CR5 3EZ'', ''KT172JW'')
14	Pcd2	8-character type of postcode(eg: ''CR5 3EZ'', ''KT17 2JW'')
15	Dointr	The date when postcode was introduced (eg: '201107')
16	Doterm	The date when postcode was terminated. This will only apply to postcodes which have been terminated and not re-used
17	Geometry	The longitude and latitude of the house

Table 1 Description of attributes in London boroughs house price dataset

Sno.	Attribute Name	Description
1	NAME	Name of the district
2	GSS_CODE	Government statistical service code.
3	HECTARES	Area of the district.
4	NONLD_AREA	Area outside London.
5	ONS_INNER	Office for national statistics code.
6	geometry	The spatial location of each of the districts
7	InOut	Depicts whether the district is part of inner borough or the outer boroughs.

Table 2 Description of attributes of London boroughs dataset

Sno	Attribute Name	Description
1	OBJECT ID	Unique ID for the schools
2	TOWN	Town name of the school
3	STATUS	Opened or Closed
4	Gender	Gender type girls/boys
5	LA_NAME	Boroughs Name
6	POSTCODE	Postcode of school
7	Geometry	Latitude and Longitude

Table 3 Description of attributes of schools dataset

S NO	Number of neighbours	Weights	Power parameter (Minkowski metric)	Number of Cross-Validation set	Model Accuracy
1	8	uniform	euclidean	4	63.4
2	14	distance	euclidean	4	66.5
3	20	distance	euclidean	5	62.66
4	50	uniform	City block	3	63.67
5	9	distance	euclidean	4	65.38
6	20	distance	euclidean	3	64.08
7	9	uniform	euclidean	4	65.57
8	90	uniform	City block	4	62.88
9	75	distance	City block	4	64.45
10	18	distance	euclidean	4	67.15

Table 4 summary table of alternative K-nearest neighbours

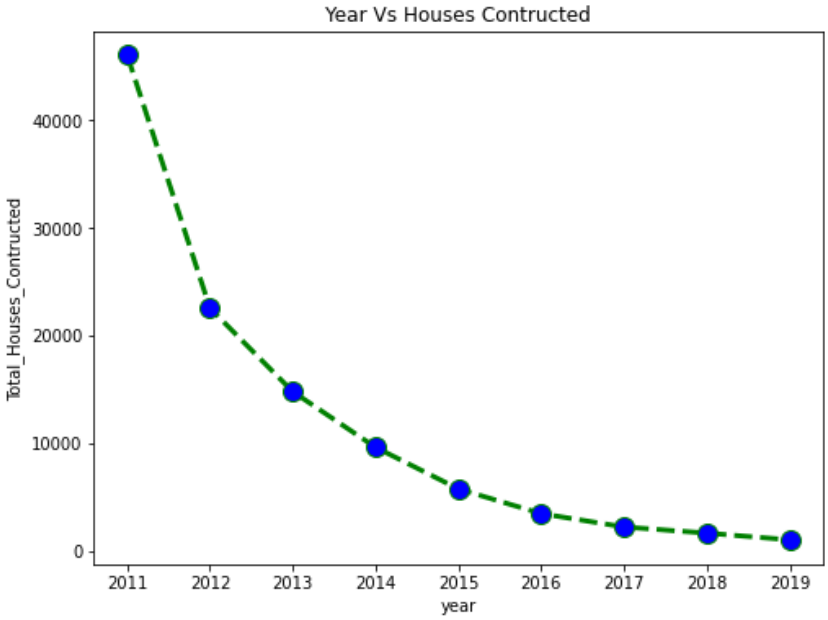


Figure 1 Construction of Houses in London throughout the Year

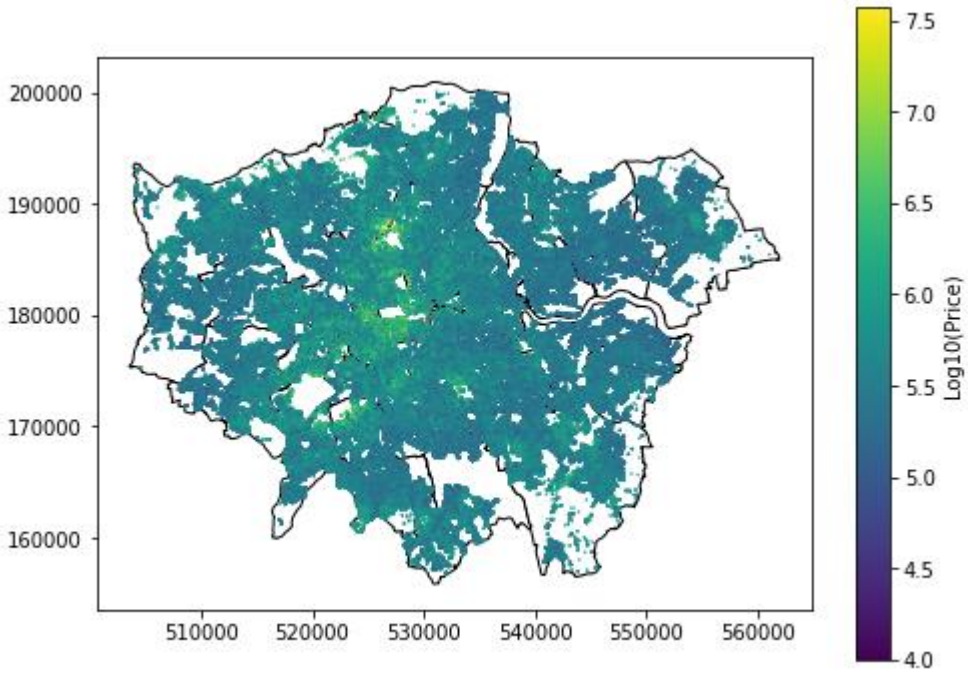


Figure 2 Scatter plot of houses of all prices (2011-2019)

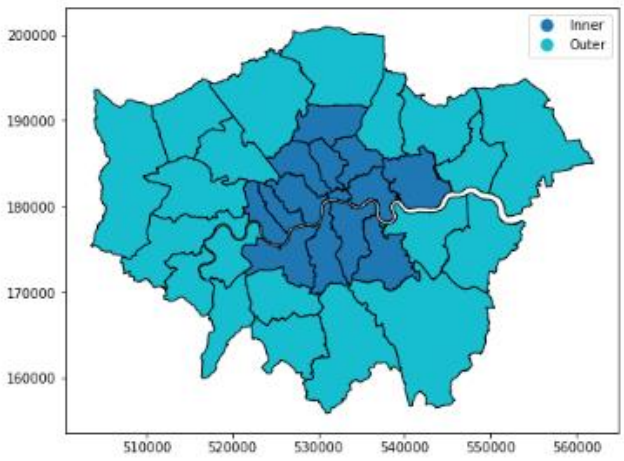


Figure 3 Inner districts vs Outer districts: This figure presents the distinction of the London borough inner region and the outer districts.

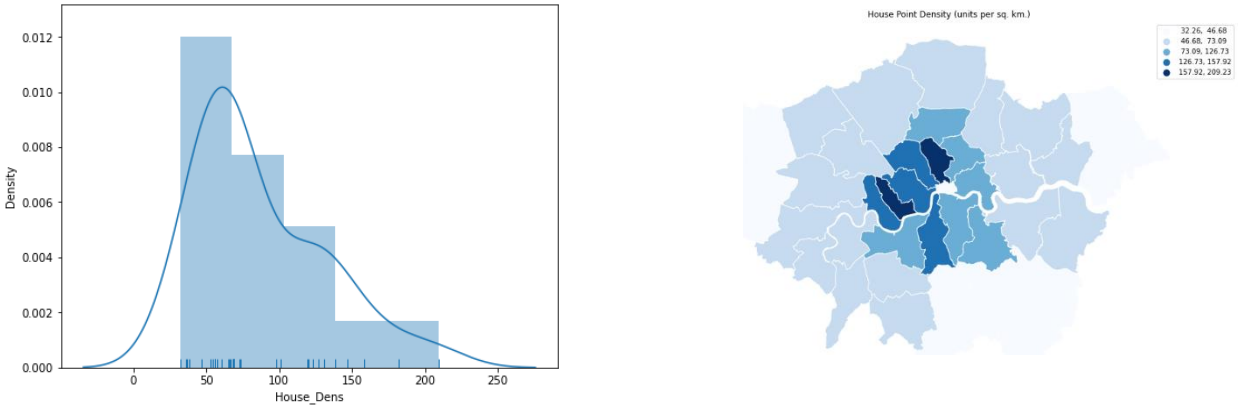
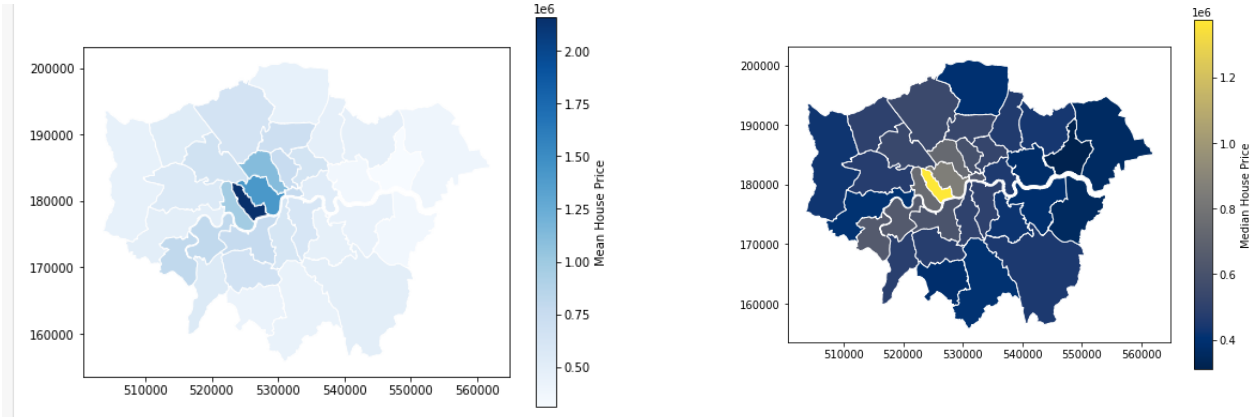


Figure 4 Density Plot for Houses in London



(a) Mean price of the districts constituting London boroughs.

b) Median price of the districts constituting London boroughs.

Figure 5 Mean price vs Median price of houses with respect to districts

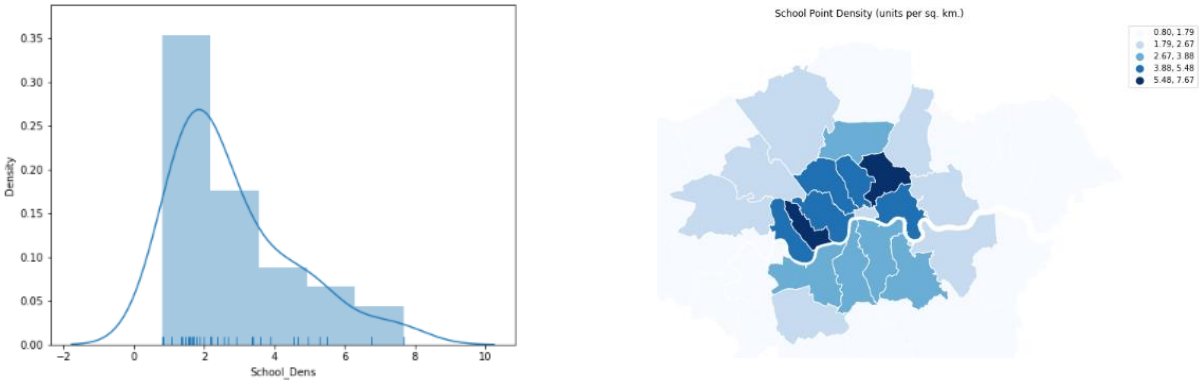
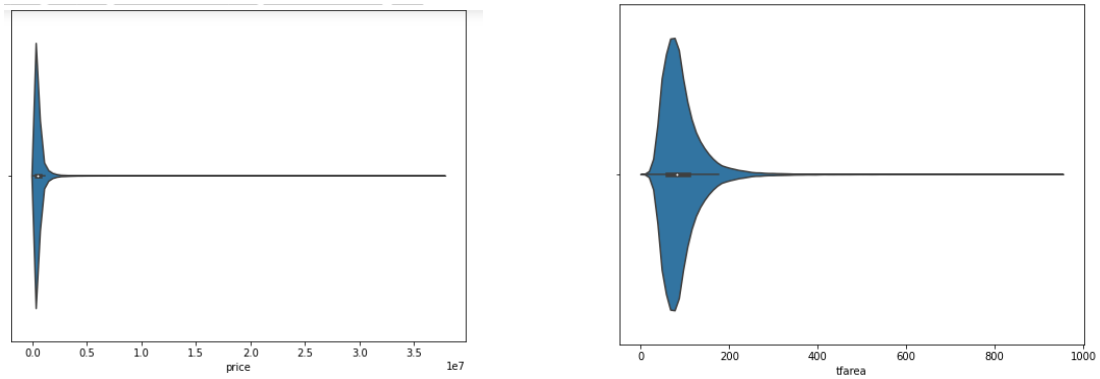


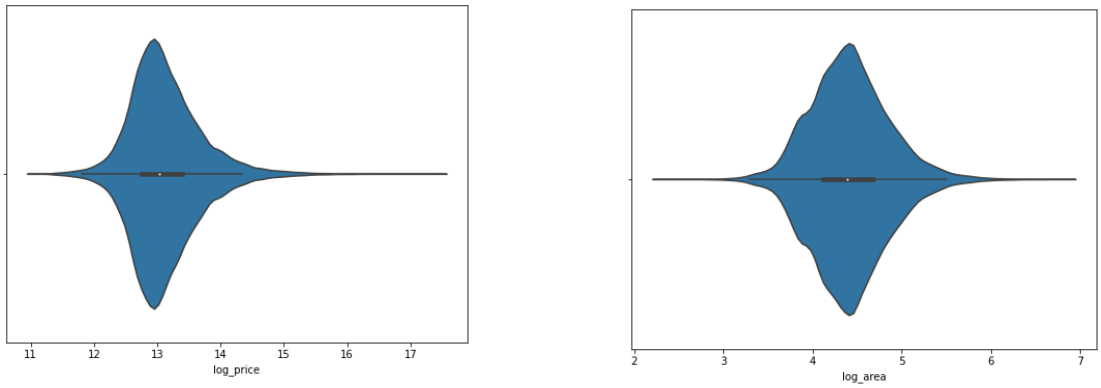
Figure 6 Density plots for Schools in London



(a) Data skewness of price attribute.

(b) Data skewness of tf_area attribute.

Figure 7 Depicts violin plot for the skewed data



(a) Normally distributed data of log(price) attribute.

(b) Normally distributed data of log(tf_area) attribute.

Figure 8 Depicts violin plot for the log transformed data

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :   log_price                Number of Observations:       59592
Mean dependent var  :     13.1134                Number of Variables   :          13
S.D. dependent var  :      0.5644                Degrees of Freedom    :       59579
R-squared           :      0.7699
Adjusted R-squared  :      0.7698
Sum squared residual:    4368.698                F-statistic           :  16608.3222
Sigma-square        :       0.073                Prob(F-statistic)     :           0
S.E. of regression  :       0.271                Log likelihood        :   -6698.757
Sigma-square ML     :       0.073                Akaike info criterion :   13423.513
S.E of regression ML:      0.2708                Schwarz criterion     :   13540.452

White Standard Errors
------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      10.9910791       0.0274388     400.5674295       0.0000000
            log_area       0.7216843       0.0067764     106.4999654       0.0000000
         numberrooms       0.0184049       0.0016673      11.0388241       0.0000000
               Flats      -0.0262610       0.0096202      -2.7297906       0.0063393
            Detached       0.1524094       0.0050574      30.1361207       0.0000000
                 New       0.1527169       0.0282491       5.4060733       0.0000001
               DEPRHH      -0.9628547       0.0086616    -111.1634602       0.0000000
            Duration       -0.0885476       0.0097263      -9.1039530       0.0000000
            Dist_KM0       -0.0000409       0.0000003    -153.0671312       0.0000000
           Dist_Road        0.0000085       0.0000010       8.2321325       0.0000000
         Dist_Transit      -0.0000179       0.0000020      -8.9404760       0.0000000
            Dist_Open      -0.0000777       0.0000080      -9.6577209       0.0000000
          Dist_School      -0.0000194       0.0000065      -2.9988324       0.0027113
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           80.217

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2       38213.395           0.0000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test               12        7704.721           0.0000
Koenker-Bassett test             12        2603.176           0.0000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Moran's I (error)              0.4009       495.047           0.0000
Lagrange Multiplier (lag)         1         301.250           0.0000
Robust LM (lag)                   1         156.010           0.0000
Lagrange Multiplier (error)       1      243339.822           0.0000
Robust LM (error)                 1      243194.582           0.0000
Lagrange Multiplier (SARMA)       2      243495.833           0.0000

================================ END OF REPORT =====================================
```

Table 5 Summary table of OLS Model

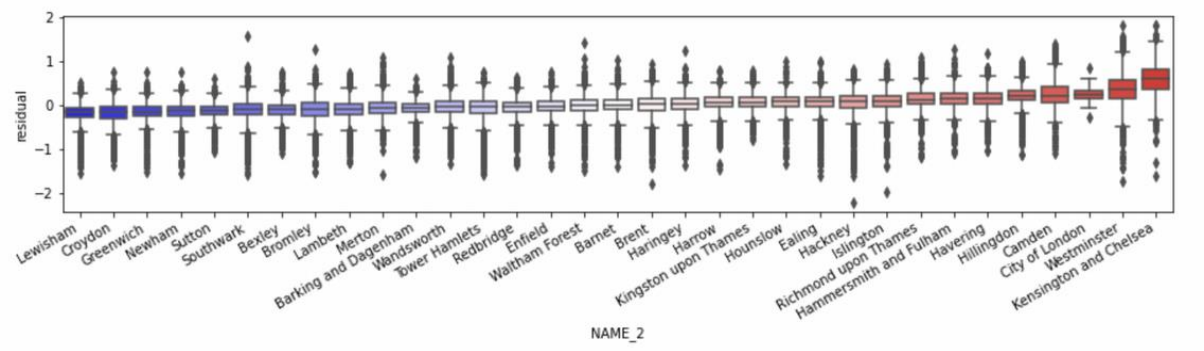


Figure 9 Represents distinction of residual prices of boroughs

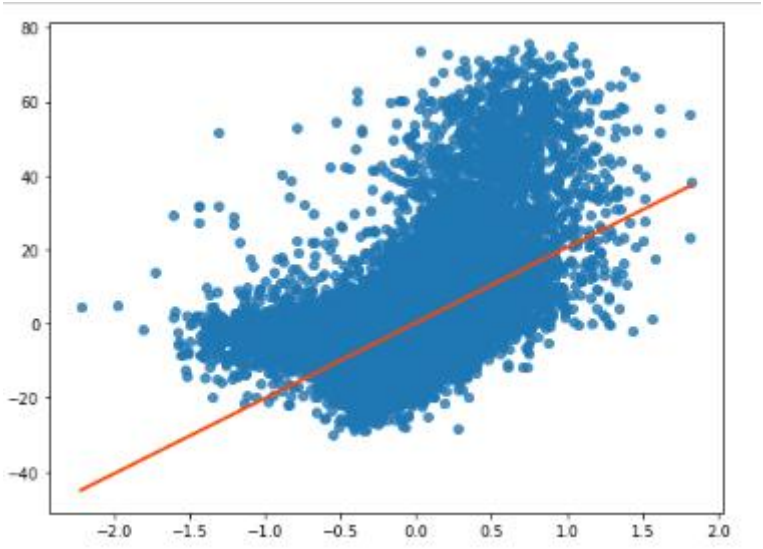


Figure 10 Represents the spatial Lag vs Model residuals

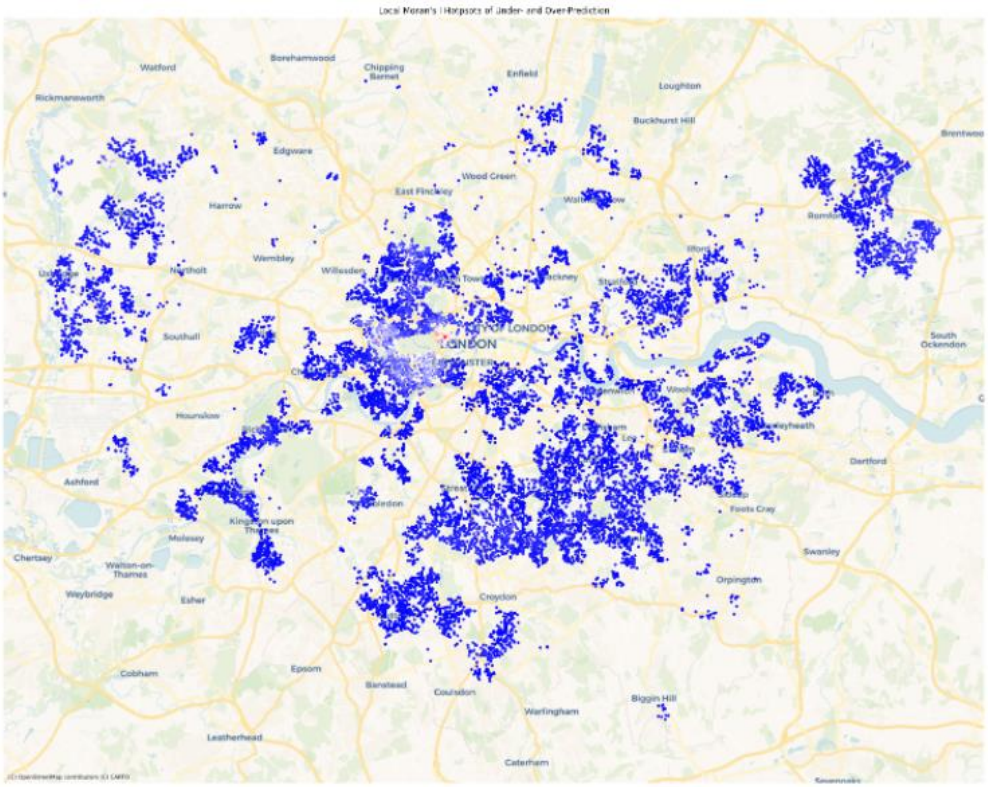


Figure 11 cluster of areas with high price and areas with low price

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES
--------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :   log_price                Number of Observations:       59592
Mean dependent var  :     13.1134                Number of Variables   :          14
S.D. dependent var  :      0.5644                Degrees of Freedom    :       59578
Pseudo R-squared    :      0.7836
Spatial Pseudo R-squared:  0.7699

White Standard Errors
------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       9.8791481       0.1625465      60.7773798       0.0000000
            log_area       0.7102968       0.0068811     103.2243644       0.0000000
         numberrooms       0.0184045       0.0016300      11.2913938       0.0000000
               Flats      -0.0298416       0.0094596      -3.1546295       0.0016070
            Detached       0.1460335       0.0052875      27.6186596       0.0000000
                 New       0.1557608       0.0274554       5.6732274       0.0000000
              DEPRHH      -0.8767982       0.0147950     -59.2633149       0.0000000
            Duration      -0.0909681       0.0096005      -9.4753810       0.0000000
            Dist_KM0      -0.0000382       0.0000005     -78.6594154       0.0000000
           Dist_Road       0.0000075       0.0000010       7.5395491       0.0000000
         Dist_Transit      -0.0000161       0.0000021      -7.7233355       0.0000000
            Dist_Open      -0.0000786       0.0000081      -9.7192729       0.0000000
          Dist_School      -0.0000059       0.0000070      -0.8376387       0.4022337
          W_log_price       0.0822120       0.0121070       6.7904380       0.0000000
------------------------------------------------------------------------------------
Instrumented: W_log_price
Instruments: W_DEPRHH, W_Detached, W_Dist_KM0, W_Dist_Open, W_Dist_Road,
             W_Dist_School, W_Dist_Transit, W_Duration, W_Flats, W_New,
             W_log_area, W_numberrooms
================================= END OF REPORT =====================================
```

Table 6 Summary table of Spatial Lag Model

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIALLY WEIGHTED LEAST SQUARES (HET)
-----------------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :   log_price                Number of Observations:       59592
Mean dependent var  :     13.1134                Number of Variables   :          13
S.D. dependent var  :      0.5644                Degrees of Freedom    :       59579
Pseudo R-squared    :      0.7555
N. of iterations    :           1                Step1c computed       :          No

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      11.2413839       0.0283456     396.5828554       0.0000000
            log_area       0.6172956       0.0053628     115.1075160       0.0000000
        numberrooms        0.0257890       0.0012720      20.2745501       0.0000000
               Flats      -0.0701168       0.0078119      -8.9756576       0.0000000
            Detached       0.1322022       0.0039868      33.1600140       0.0000000
                 New       0.1197068       0.0234958       5.0948181       0.0000003
              DEPRHH      -0.5623492       0.0103634     -54.2628133       0.0000000
            Duration      -0.1425922       0.0079334     -17.9735564       0.0000000
            Dist_KM0      -0.0000448       0.0000014     -31.5535306       0.0000000
           Dist_Road       0.0000213       0.0000048       4.4027630       0.0000107
         Dist_Transit      -0.0000091       0.0000064      -1.4322744       0.1520653
            Dist_Open       0.0000020       0.0000092       0.2211867       0.8249471
          Dist_School       0.0000332       0.0000080       4.1613761       0.0000316
               lambda       0.8698339       0.0040606     214.2113814       0.0000000
------------------------------------------------------------------------------------
================================= END OF REPORT ====================================
```

Table 7 Summary table of Spatial Error Model

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIALLY WEIGHTED TWO STAGE LEAST SQUARES (HET)
-----------------------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :   log_price                Number of Observations:       59592
Mean dependent var  :     13.1134                Number of Variables   :          14
S.D. dependent var  :      0.5644                Degrees of Freedom    :       59578
Pseudo R-squared    :      0.7566
Spatial Pseudo R-squared:  0.7563
N. of iterations    :           1                Step1c computed       :          No

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      11.2101865       0.0802566     139.6792341       0.0000000
            log_area       0.6208617       0.0054581     113.7512749       0.0000000
         numberrooms       0.0254351       0.0012736      19.9710448       0.0000000
               Flats      -0.0692314       0.0078097      -8.8648334       0.0000000
            Detached       0.1327210       0.0039801      33.3462762       0.0000000
                 New       0.1200264       0.0234465       5.1191529       0.0000003
              DEPRHH      -0.5702680       0.0104311     -54.6700106       0.0000000
            Duration      -0.1412916       0.0079315     -17.8140675       0.0000000
            Dist_KM0      -0.0000446       0.0000014     -32.4932283       0.0000000
           Dist_Road       0.0000198       0.0000050       3.9391620       0.0000818
        Dist_Transit      -0.0000102       0.0000066      -1.5573907       0.1193778
           Dist_Open       0.0000014       0.0000092       0.1491051       0.8814707
         Dist_School       0.0000337       0.0000077       4.3535448       0.0000134
         W_log_price       0.0014475       0.0053405       0.2710467       0.7863551
              lambda       0.8772844       0.0045357     193.4168613       0.0000000
------------------------------------------------------------------------------------
Instrumented: W_log_price
Instruments: W_DEPRHH, W_Detached, W_Dist_KM0, W_Dist_Open, W_Dist_Road,
             W_Dist_School, W_Dist_Transit, W_Duration, W_Flats, W_New,
             W_log_area, W_numberrooms
================================= END OF REPORT =====================================
```

Table 8 Summary table of Spatial Lag + Error Model

REFERENCES

Black, S. (1999). Do Better Schools Matter? Parental Valuation of Elementary Education. *The Quarterly Journal of Economics*, Vol.114, 577-599.

developers, p. (2019). *spreg documentation.* Retrieved from https://spreg.readthedocs.io/_/downloads/en/master/pdf/

Government, U. (n.d.). *GOV.UK*. Retrieved from HM Land Registry: https://www.gov.uk/government/publications/about-the-uk-house-price-index/about-the-uk-house-price-index

Stephen Gibbons, S. M. (2012). Valuing school quality using boundary discontinuity. *Journal of Urban Economics*.

Yinger, J. &.-H. (2011). The Capitalization of School Quality into House Values. *Journal of Housing Economics*.