Exploring and Analysing factors which affect the Housing Price.

Snehal Deshmukh, Shreya Agarwal, Madhusudan Panwar

Contents:

Introduction, Research question, Background:

This research looks at the level of London's borough house rates from a money perspective.
It looks at the basic procedures that manages the trade for buying and selling homes, considering the area to which the observed swaps in house rates are sensible, and feasible long-term results of house price inflation for London.
For predicting house prices, it needs more precise method based on location, house type, size, build year, transportations and some other components which could influence house demand and supply. In fruitful terms, house prices in stability are placed by the balance of demand and supply.
Housing is also, in many respects, a good that is demanded indirectly – in terms of access to local facilities, employment opportunities and other services it provides [1]. Due to this, the preference to buy a home will be affected by demand for these other markets. High house prices in a particular area may therefore reflect a relative abundance of amenities and offer residents a high quality of life [2].
Spatial location widely focus to examine the role of geography and area in economic phenomena, and a particular strand of research is devoted to the analysis of real property market variations as one of the economic situations in a particular geographic area.

The general idea of our project is analysing the house prices on the basis of some interesting, concerning factors. We tend to check if these factors are influencing the increase or decrease in housing prices. If specified in an explanatory way, we are looking at "what" characteristics of the house and also neighbourhood, are affecting the prices and how they are changing according to the rate of change in these, more significant factors.
Hence, hypothesis taken is, - "Are the housing prices dependent on the considered explanatory variables?" where,

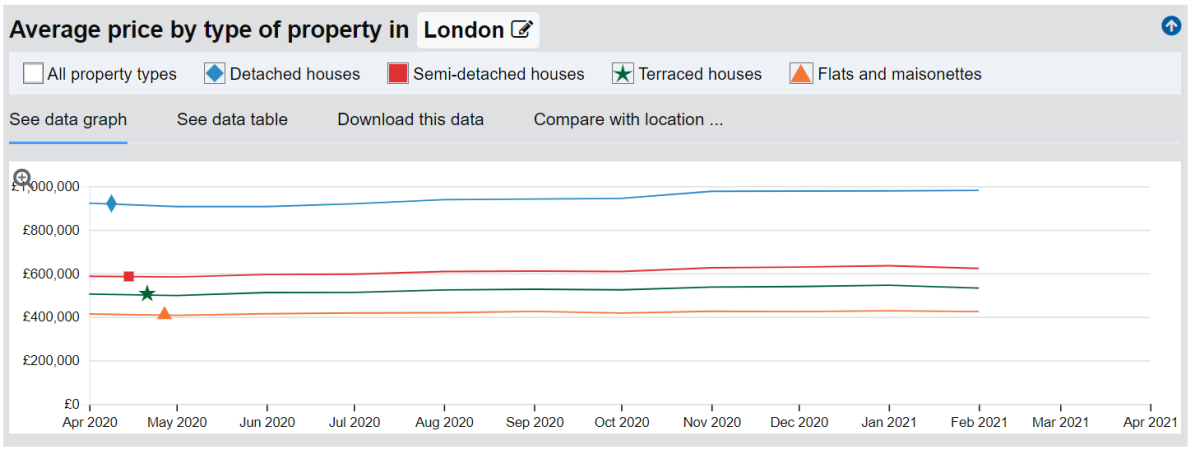
> Null hypothsis(H0): There is no dependency of these factors/variables on house prices.
> Alternative hypothesis(Ha): There is less/more dependency of these factors/variables on house prices.

To be more specific towards the research question, "Does the 'Property type', 'Duration-freehold or leasehold', and 'distance from primary roads and transit stations' change the house pricing? And if yes, How, to what extent these affect the change?" i.e., the percentage change in house prices with respect to the above-mentioned factors.
Conceptual significance: From money perspective, it looks at the basic procedures that manages the trade for buying and selling homes.
Practical significance: Considering the area to which the observed swaps in house rates are sensible, and feasible long term results of house price inflation for London.

We are considering the Hedonic framework which takes into account, the neighbourhood effects (such as distance of the house from the stations, roads, and also from the city centre) along with the characteristics of the house (including numberofrooms, property type, etc). This is done by analyzing spatial data using various geovisualisation techniques using methods of Quantitative data classification, color theory, Areal data considerations, then studying global spatial autocorrelation using concepts of Spatial lag, Global Moran's I, spatial indications in PySal and at last, checking and fitting models using OLS regression, spatial regression, Spatial feature Engineering, SAR, SEM, SARMA and Spatial Regimes model in SPREG.

Our data study and analysis mainly focus on factors- Property type, Duration, and Distance/Location of the House. What these factors are? Let's have a deeper understanding of them.

Property type is the type of house/structure a consumer will look or buy according to his interest. These types in our dataset include Detached, Semi-Detached, Terraced, Flats/Maisonettes, and Other. In general, detached type of house should have higher rates as it is an independent structure and gives the owner his/her own space. For London, normally, the "detached" type of house is expensive. As we researched deeper on this particular variable, we came across some interesting facts. For London, we saw that detached type of houses were actually expensive than the other types, hence our reason to choose this particular type in analysing "how" it affects the house prices. Figure showing "the track of average prices for all property types or focus on one in particular" [4] for London for year 2020 and 2021 overall, is as given below. Figure.[4]



Then, another important characteristic we came across is "Duration". It shows if the property is for freehold or leasehold. We chose to see if the freehold duration type is significant in looking at the house prices. In practice, a residential freehold interest applies to the outright ownership of land or property for an unlimited period.
After studying, we saw that - Freehold can also be shared, where the freehold of the building is (a) either owned jointly by a number (up to four) of the flat owners in their personal names, or (b) where a company is the owner of the freehold and each of the leaseholders hold a share or membership in that company [5].
A freehold property, hence, is any real estate that is legally 'free from hold' of any entity other than the owner. The owner of such a property has the right to use it for any purpose, in accordance with the regulations of where it is located. The sale of a freehold property requires significantly lesser paperwork, as it is not necessary to request authorisation from the state. However, this also means that a freehold property is more expensive to purchase than a leasehold property.[6]

House prices are affected by distance, transportation services, local amenities, and many other neighbourhood effects to important economic centres that provides many employment opportunities and an extensive assistance network.
The economic geography writings emphasizes the significance of mobility, transportation costs, and travel time for the extension of the borough as it is crucial in deciding where to purchase a house because of shorter distance or less travel time implicit more free time, therefore one might be ready to pay a higher amount for such type of houses.
These are the geographical components that cannot be ignored as driving forces for price diversity.

The worth of the locality is actually related to its availability and reach to economic centres.
Hence, we with the characteristic features of the house, we have also taken into account more spatial features such as distance of the property from city centre, from transit stations and from the primary roads. Most significant consideration overall being the distance from city centre as justified above, in the previous paragraph. Also, we will be looking at other two factors more deeply further as we proceed, analysing them according to each Borough (Borough level analysis).

Data, Study area, and Methods:

Data:

London Housing dataset of around 9 years (2011 to 2019) having basic features of house (tfarea, numbrooms etc) having a geometry column (point dataset) which means every house is represented as point in this geojson dataset.
London Boroughs dataset having division of London into 33 small local administration, every borough is represented as multipoint polygon in this geojson london borough dataset, both of which were in EPSG 27700 and converted to EPSG 3857.
For the 3rd part, the datasets used- "data_all 2014", "KM_Zero", "OutputAreas" and "london_boroughs". Also used "houses_lr", "OpenSpace", "Primary_Roads" and "TransitStations" which is combined into one single geojson file-"data_all_2014" for year, 2014.
The open space, roads and transit station distance variables are added.

Study area:

First, we found all the boroughs in which all houses are located. Average cost per unit area in each and every Borough to find out which boroughs are costly and what all factors are there that will affect the price of borough.
Next, the centre of area in this part is “Exploratory Spatial Data Analysis in PySAL” where we aim to recognize where these high and low outliers are situated and dived to analyse and understand correlation that how are they spatially autocorrelated.
Finally, the focus of area was "Explanatory spatial regression models using spreg". In a simple way, it is, what we do with the data when we have found something interesting about it, then want to focus on what exactly has happened and how did it happen. It is related to gathering information and then fitting different models to produce results which will in turn give new insights and knowledge. For example, in our case, we will be performing some regressions on the data in order to find out how the variables are helpful in predicting the house prices. It will tell that by how much % change in them will change the house price (and vice versa).
We study using Spatial Regression models, used to model the spatial relationships between the predictors for decision making.

Methods:

Standardization - Raw counts of aggregated areal data are not directly comparable so we did standardization by taking house density into consideration rather than house count.

Spatial concentration - There is error associated with the data attached to areal unit, because it is often hard to get a spatially-representative sample of the individuals within that unit so we looked at granular level instead of simply taking borough division.

Quantitative data classification using mapping:
Equal Interval: The basic concept of the equal interval scheme is that each bin contains an equal width (w) of the attribute value for a specified number of bins (k). In mapclassify we can specify our preferred k value, and the EqualInterval function automatically divides the variable of interest into k bins.
Quantiles: An alternative approach is to create bins of equal numbers of observations (n), rather than width, by dividing n by k and placing the breaks sequentially from the minimum to maximum value. Thus, for k= 5, the first bin contains the smallest 20% of data values, while the last bin contains the largest 20% of data values.
Mean-Standard Deviation: We are interested in better understanding the location and context of outliers, which can be done using the mean-standard deviation classifier. This scheme defines class boundaries as some distance from the attribute mean in terms of multiples of the standard deviation of the attribute.
Fisher-Jenks: This is representative of a number of similar schemes that use a heuristic approach to optimize the breaks between bins by attempting to minimize the sum of absolute deviations around class means.

Then, using Exploratory Data Analysis-
1. Global Spatial Autocorrelation: To determine the extent to which I's values of X are correlated with the values of X in I's neighbourhood.
According to Tobler's first law, everything is related to everything else, but near things are more related than distant things which is also known as spatial similarity, where we used spatial lag, join-count, Global Moran's I for assessing the correlation.
The first step in computing these spatial associations we designed the spatial weights matrix, which determined the nearness using the most common type of weight is queen contiguity weight, which reflects adjacency relationships through a binary indicator variable whether or not a polygon shares an edge or a vertex with the other polygon.
Spatial lag seizes the behaviour of a variable in the immediate neighbouring of each location which is related to local smoothing of a variable which estimates the value on the basis of an average of other observations which will be termed as "close" with respect to one or a set of explanatory variables which also reduce noise.
Join count is most frequent type for assessing the correlation. A join exists for each adjacent pair of observations in terms of a binary variable (0=white and 1=black). Eg., black over £6,500 per square meter.
Join count test examine this in a very simple way by tallying up the number of different types of joins: black-black, white-white and black-white these are observed observations and comparing these "observed" to null hypothesis "expected" number of counts based on the concept of complete spatial randomness where black and white cells are randomly assigned across the region. Calculating the number of black-black occurrences, white-white occurrences, black-white occurrences (observed) with those of CSR expected values.
Global Moran's I looks at the correlation between a given unit's value of X and the average value of X in its neighbours. To understand more clearly, Moran scatterplot can be constructed which simply depicts the relationship between X and lag of X.
If needed variables are standardised by taking z-scores so that they are represented in units of standard deviation from their respective means.

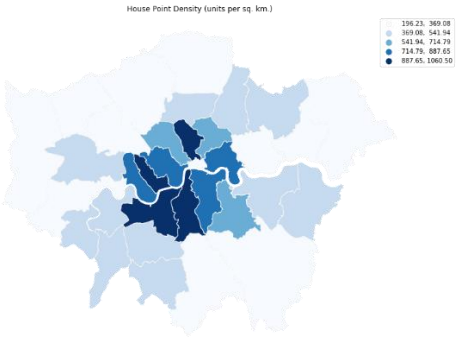
2. Local Indicators of Spatial Association (LISA) measures local autocorrelation. They are used to test for significance by testing whether the correlation between an observation and the average of its neighbours i.e., the local Moran's I is more similar (high- high, low- low) or dissimilar (high-low, low-high).

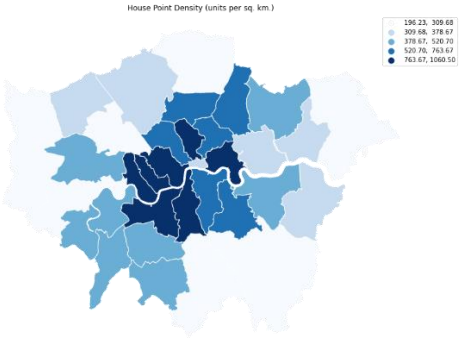
Lastly, we are referring various explanatory spatial regression methods using spreg package in python to analyse the housing prices in a hedonic framework (using the Hedonic regression model which are used to explain the effect of various characteristics on urban housing values).
First, we applied Non-spatial OLS regression in order to study the explanatory variables (House characteristics, Eg. Number of rooms, Total floor area etc).
We headed further by plotting Moran's I map while including information about the neighbourhood i.e., forming spatial clusters of residuals.
Later, after checking the neighbourhood effects, we calculated the distance of each house from the city centre using geometry.apply() and again mapped them.
We added more spatial features such as distance from primary roads (dist_roads) and trees density (TreeDens) and fitted the OLS models every time while adding new features using SPREG.
Then, we checked for the best model to use while considering all these features. Models fitted on this data were Spatial Lag Model (SAR using GM_Lag): considers dependent variables on an area with other areas associated with it[1], Spatial Error Model (SEM using GM_Error_Het): takes into account the dependency of error values of an area with errors in other areas associated with it[3], and Spatial Lag + Error Model (SARMA using GM_Combo_Het- compares both lag and error components). At the end, according to the model best fitted, we applied it to spatial regimes model (which performs individual regression for each regime to explain the neighborhood-level heterogeneity).

Results:

The analysis started with the concept of geovisualization in the context of choropleth maps for areal data on London boroughs dataset on social/demographic data and a sample of the London house price dataset on housing prices and characteristics. The main focused areas- **quantitative data classification** and **color theory** for choropleth maps.
1st was quantitative data classification which can be done in 4 ways- a) Equal intervals b) Quantiles c) Mean-standard deviation d) Fisher-Jenks as given below-

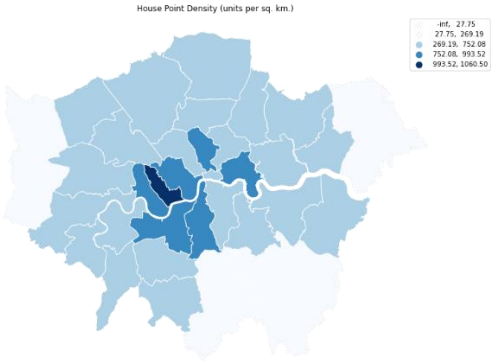
a) b)

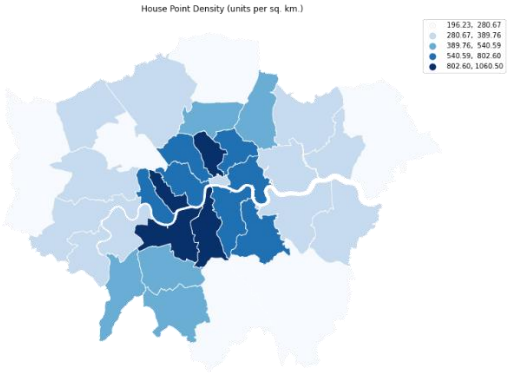




c)

d)

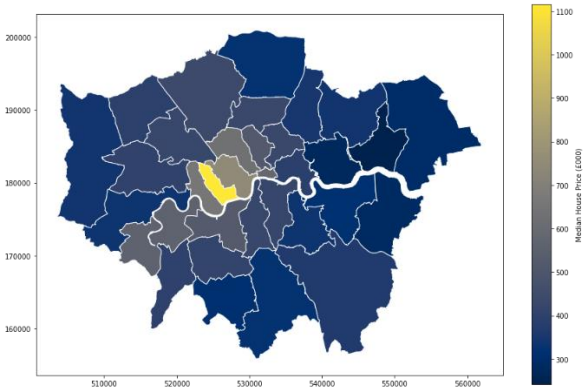


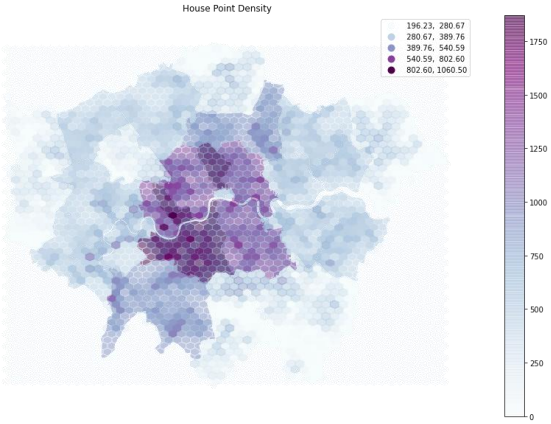


2nd, in Color Theory, a) Below is a plot for median price for each borough and colouring by price, telling us that distance matters while predicting the prices, which can be seen by different color hues as below.

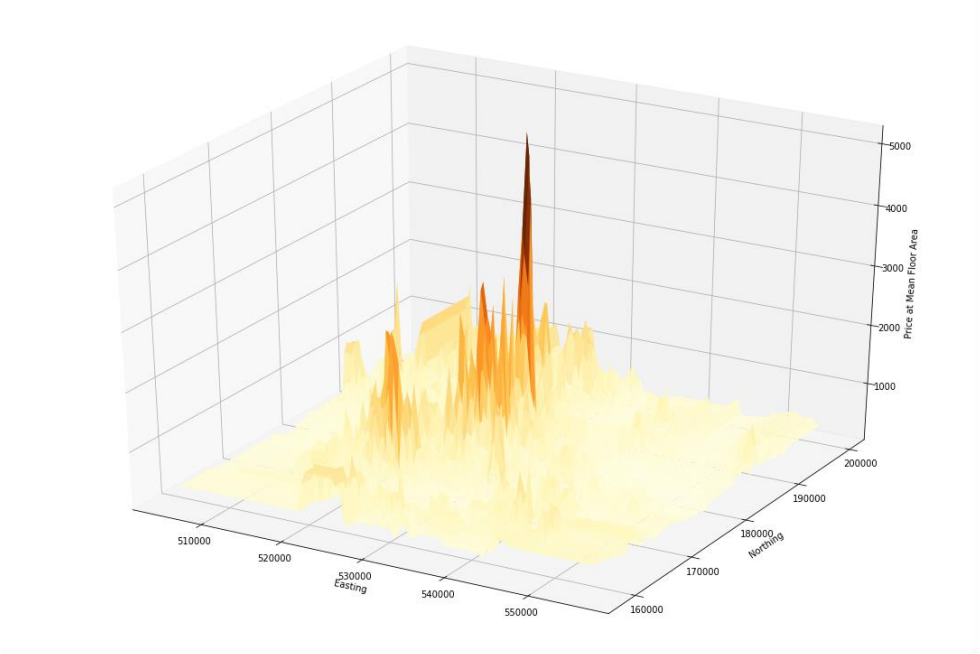
b) Spatial Concentration showed us how minute we can go on visualising the data to a specific point.

a) b)

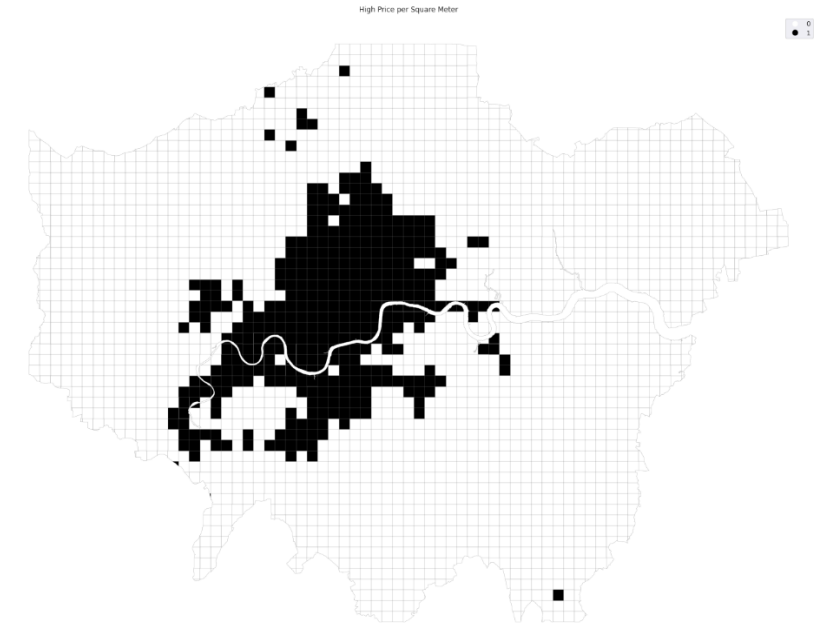




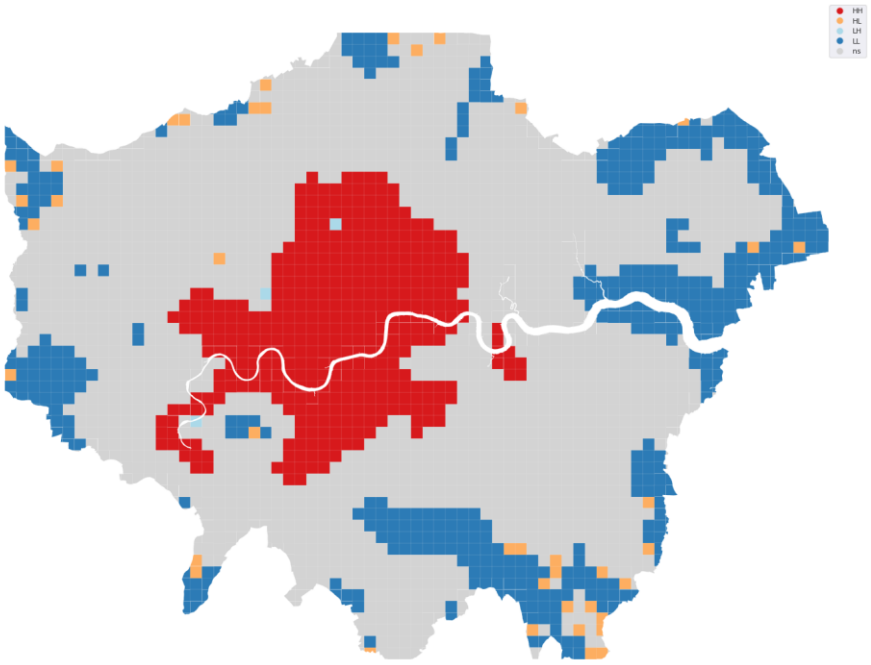
We also performed, KNN to predict house prices, map (it shows Price of Mean floor area versus Northings and Eastings of the house) of which is as below-



Heading towards the exploratory Data Analysis, from the map there seems to be a clear pattern of spatial autocorrelation: black cells tend to be located near other black cells, and white cells near one another (i.e., there is a stark cluster of cells with high values of median price per square meter near the London city centre; black tiles specify the high values of median house per square meter).



Another thing was, we were able to identify "hostpots" and "coldspots" within sub-regions of observations. These local spatial autocorrelation statistics tells us the whether the correlation between an observation and the average of its neighbours is more similar (high-high, low-low) or dissimilar (high-low, low-high) than we would expect from pure chance.
Looking at it and then mapping using LISA, we concluding that there was high correlation between the houses which were somewhere near to the city centre and between ones which were situated in the outer areas as can be seen in the plot below-



Coming to answer the above stated research questions, we performed explanatory spatial regression considering distance of 1km in distance-based spatial weights matrix and the house characteristics as- Terraced, detached types of property. Later we considered neighbourhood and spatial effects i.e., distance from city centre, from primary roads (dist_Road) and from transit stations (dist_Transit).
The **results** are as followed-

1) Housing prices are dependent on “Detached” type of property by exponential of (15%)(As we have the price taken in log), as we see from the below table. Also, considering our hypothesis, we see that the p-value (<<0.01) is very low justifying us to reject the Null hypothesis.
Overall, we can expect a 16% increase in housing price if it is of detached type.

2) Similarly, for “Freehold”, there is its dependency on house prices.
We can expect a 17.06% increase in house price if it is “Freehold”. Again, we reject null hypothesis here (p-value is very less).

3) Significance of distance from primary roads showed some level of dependency (though very little).

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIALLY WEIGHTED TWO STAGE LEAST SQUARES (HET)
-----------------------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :   log_price                Number of Observations:       90686
Mean dependent var  :     12.8852                Number of Variables   :          13
S.D. dependent var  :      0.6458                Degrees of Freedom    :       90673
Pseudo R-squared    :      0.7562
Spatial Pseudo R-squared:  0.7538
N. of iterations    :           1                Step1c computed       :          No

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      10.6961671       0.0250959     426.2122280       0.0000000
            log_area       0.7487454       0.0061647     121.4573008       0.0000000
         numberrooms       0.0180957       0.0015767      11.4772268       0.0000000
             Terrace      -0.0828677       0.0030920     -26.8004263       0.0000000
            Detached       0.1527052       0.0051766      29.4993341       0.0000000
                 New       0.0874124       0.0139242       6.2777174       0.0000000
            Freehold       0.1575573       0.0041932      37.5742741       0.0000000
            Dist_0KM      -0.0000492       0.0000003    -159.7027808       0.0000000
           Dist_Road       0.0000113       0.0000010      11.0999865       0.0000000
        Dist_Transit      -0.0000153       0.0000020      -7.5462306       0.0000000
      Dist_OpenSpace      -0.0001109       0.0000077     -14.4035985       0.0000000
              DEPRHH      -1.2363267       0.0083724    -147.6671159       0.0000000
         W_log_price       0.0000374       0.0000011      34.7089271       0.0000000
              lambda       0.0000000       0.0000043       0.0000000       1.0000000
              lambda       0.0000000       0.0000043       0.0000000       1.0000000
------------------------------------------------------------------------------------
Instrumented: W_log_price
Instruments: W_DEPRHH, W_Detached, W_Dist_0KM, W_Dist_OpenSpace,
             W_Dist_Road, W_Dist_Transit, W_Freehold, W_New, W_Terrace,
             W_log_area, W_numberrooms
================================= END OF REPORT ====================================
```

4) From the results, we saw that the spatial lag model was more appropriate for this data considerations. Hence, after fitting it on individual Borough level, we found that for “Kensington and Chelsea”, Distance from primary roads and transit stations, both showed some dependency whereas for “Sutton”, only major influencing factor was distance from roads.
More discussion on this is done in the following parts of the paper.

```
hpken = hp_0km_3[(hp_0km_3["NAME_2"]=="Kensington and Chelsea")]
wken = weights.DistanceBand.from_dataframe(hpken, 1000)
m7ken = spreg.GM_Lag(hpken[['log_price']].values, hpken[all_names].values,
                     w=wken, name_y='log_price', name_x=all_names)
print(m7ken.summary)
```

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES
--------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :   log_price                Number of Observations:        1924
Mean dependent var  :     14.0955                Number of Variables   :          13
S.D. dependent var  :      0.8575                Degrees of Freedom    :        1911
Pseudo R-squared    :      0.8459
Spatial Pseudo R-squared:  0.8459

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      10.5263314       0.1243705      84.6368807       0.0000000
            log_area       0.9979042       0.0258216      38.6461631       0.0000000
        numberrooms        0.0214124       0.0077845       2.7506339       0.0059480
             Terrace       -0.0428073       0.0498538      -0.8586579       0.3905293
            Detached       -0.0303428       0.1096846      -0.2766369       0.7820589
                 New       -0.1392458       0.0847234      -1.6435346       0.1002724
            Freehold        0.2067522       0.0520820       3.9697417       0.0000720
            Dist_0KM       -0.0001573       0.0000088     -17.9109267       0.0000000
           Dist_Road        0.0001082       0.0000274       3.9541194       0.0000768
       Dist_Transit        -0.0001074       0.0000352      -3.0488979       0.0022968
     Dist_OpenSpace        0.0001110       0.0000804       1.3798967       0.1676185
             DEPRHH        -0.6411349       0.0616816     -10.3942667       0.0000000
        W_log_price        0.0000098       0.0000039       2.5009743       0.0123852
------------------------------------------------------------------------------------
Instrumented: W_log_price
Instruments: W_DEPRHH, W_Detached, W_Dist_0KM, W_Dist_OpenSpace,
             W_Dist_Road, W_Dist_Transit, W_Freehold, W_New, W_Terrace,
             W_log_area, W_numberrooms
================================ END OF REPORT =====================================
```

```
hpsu = hp_0km_3[(hp_0km_3["NAME_2"]=="Sutton")]
wsu = weights.DistanceBand.from_dataframe(hpsu, 1000)
m7su = spreg.GM_Lag(hpsu[['log_price']].values, hpsu[all_names].values,
                    w=wsu, name_y='log_price', name_x=all_names)
print(m7su.summary)
```

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES
--------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :   log_price                Number of Observations:        2933
Mean dependent var  :     12.5974                Number of Variables   :          13
S.D. dependent var  :      0.4287                Degrees of Freedom    :        2920
Pseudo R-squared    :      0.8587
Spatial Pseudo R-squared:  0.8587

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      10.2382207       0.0777753     131.6383796       0.0000000
            log_area       0.4443665       0.0166108      26.7516228       0.0000000
        numberrooms        0.0463540       0.0042330      10.9505633       0.0000000
             Terrace       -0.0419041       0.0087741      -4.7758848       0.0000018
            Detached        0.1418506       0.0127491      11.1263604       0.0000000
                 New        0.0638915       0.0448804       1.4235947       0.1545638
            Freehold        0.2421798       0.0109783      22.0598901       0.0000000
            Dist_0KM        0.0000263       0.0000030       8.8934771       0.0000000
           Dist_Road       -0.0000207       0.0000065      -3.2011192       0.0013689
       Dist_Transit        0.0000085       0.0000087       0.0527453       0.9579348
     Dist_OpenSpace        0.0000918       0.0000263       3.4957146       0.0004728
             DEPRHH       -0.5748414       0.0324597     -17.7093786       0.0000000
        W_log_price       -0.0000068       0.0000024      -2.8124963       0.0049159
------------------------------------------------------------------------------------
Instrumented: W_log_price
Instruments: W_DEPRHH, W_Detached, W_Dist_0KM, W_Dist_OpenSpace,
             W_Dist_Road, W_Dist_Transit, W_Freehold, W_New, W_Terrace,
             W_log_area, W_numberrooms
================================ END OF REPORT =====================================
```

The factual proofs on the basis of our study, along with the already studied (more common) house characteristics, confirm that the other most dominant factors for driving the value of a house are the property type, if it is detached or not, occupies freehold or leasehold and location (distance from city centre, primary roads and transit stations): by showing how distinct locations have a powerful smash on their prices.

Discussion and Conclusion:

Kensington and Chelsea is a Borough near to city centre in which the house prices are high because of good access to local facilities, employment facilities and other various services it provides, considering all the amenities nearby. Taking all these factors into consideration, it makes more sense that the preference to buy a home will be affected by demand for these locations in the market. High house price in a particular area offer residents a high quality of life. Moreover, while purchasing a house in city centres with good feasibility. But, in this case, another affecting factor is distance to primary roads even if the Borough lies in city centre. People tend to consider the transportation costs and might prefer less travel time with shorter distance implicit more free time, one might be ready to pay a higher amount for such type of houses. The accessibility to primary roads even after being in city centre, might be beneficial so as to travel across the country or region, i.e., to travel out of the city along with within the city. Also, we saw little dependency on distance from transit stations. This Borough is Royal and the number of people has dramatically increased. As the stations are more accessible but the place can be very crowded and polluted in comparison to the areas which are atleast little far from the stations, people prefer to buy houses which are distant from them. Hence, far is the station in city centre from the house, more expensive is the house.

Sutton is a Borough in outer area of London, hence it might be better if the location of the house is near the open space rather than near the roads as anyhow, people will have to travel to reach out to other areas as they are in outskirts, which means people there might be prefer more greenery and serenity over roads nearby. So, we can predict that, less number of roads are near the house, more is the price.

The point of perspective presented is different interrelated with respect to the writings on the field, which examined attributes achieving the type of building, location and its neighbourhood effects. This is the main motivation that our dataset has been selected.
This report investigated the fluctuations in London housing price of year 2014. Through spatial analysis using spatial feature engineering, spatial diagnostic (spatial lag model, spatial error model, spatial lag+ error model and spatial regimes model) of their possible relationship, spatial model of housing prices is built.
House of type detached, distance from the city centre, freehold property type, distance from primary roads and transit stations are the most significant factors that influence the house price variations in London boroughs which indicates that accessibility, location, property type are predominant point cuts to force housing increase.
Therefore, we can say that house price in London near city centre are high, and growing so. There is anyway, jumbled proof of whether these are overestimated or when reviewed based on other metrics. Hereafter direction of analysis may examine taking in extra property transaction data from a larger geographical area with more features, or analysing other types of property at a larger level of counties.

Reference:

[1] House prices in London – an economic analysis of London's housing market (November 2015): Joel Marsden

[2] Real Earnings Disparities in Britain (January 2011): Stephen Gibbons (SERC, Department of Geography & Environment, London School of Economics) Henry G. Overman (SERC, Department of Geography & Environment, London School of Economics) Guilherme Resende (SERC, Department of Geography & Environment, London School of Economics)

[3] Spatial autoregressive with a spatial autoregressive error term model and its parameter estimation with two-stage generalized spatial least square procedure: *D R S Saputro1 , R Y Muhsinin2 , P Widyaningsih3, and Sulistyaningsih4*

[4] UK House Price Index (data.gov.uk)

[5] Ministry of Housing, Communities & Local Government: Estimating the number of leasehold dwellings in England, 2018-19

[6] What is freehold property: Meaning, benefits and owner's rights (housing.com)