Housing price estimation in London

Ruben Campos

15/10/2022

Contents

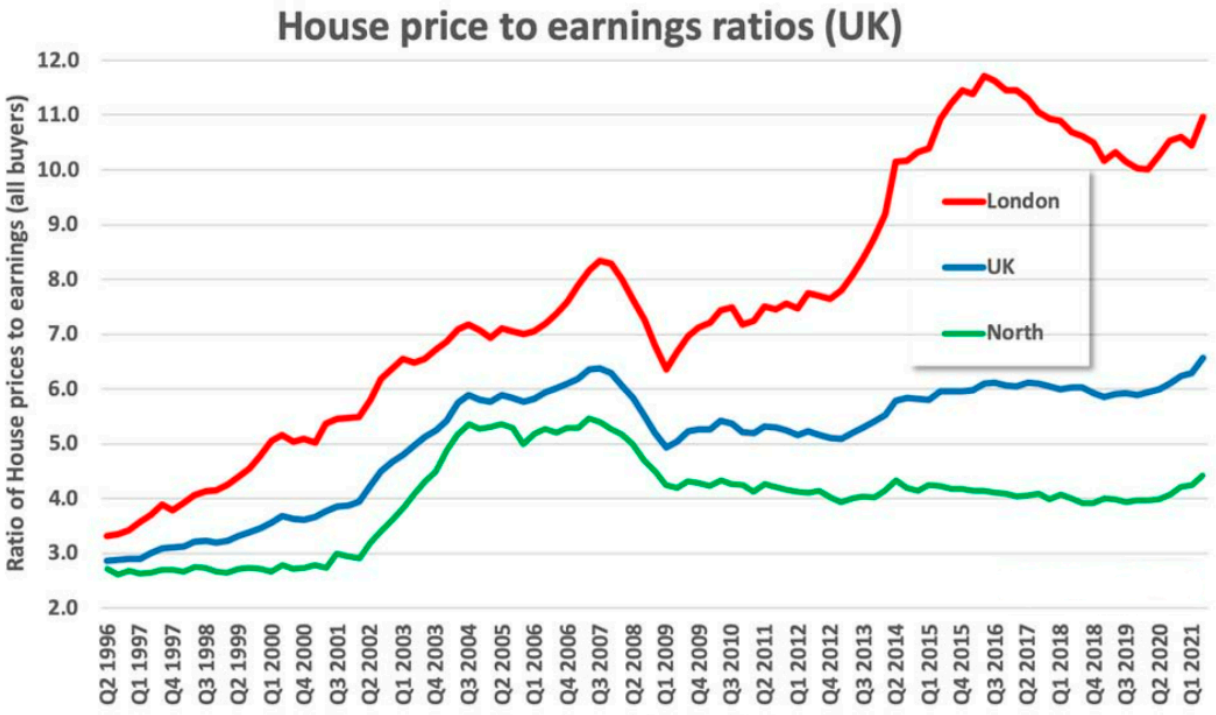
1. Project Introduction

I lived in London for seven years, as a teenager. It was the late 1980s and early 1990s and house rental or even purchase prices, while not low, were not as higher as they are right now. Checking the official data, we can see that in 1995 the average price of a house in London was 74.721£; in the last price revision, in July 2022, the official statement *London Data Store* indicated that the average house price in London of a house was 527.491£. In fact, London, together with Geneva and Zurich, has the highest average cost of an apartment in Europe in the 1st quarter 2022: 13.593,52£ per square meter.

There are many factors to understand this price raid, and we are going to detail them briefly, although the objective of this project is, taking into account the data provided by government authorities, to establish the strong relationships between house prices in the London metropolitan area and different parameters such as the floor area, the location, number of rooms or bathrooms… and the price that a home can reach according to the relevance of these parameters.

2. Understanding a problem: the increase in housing prices

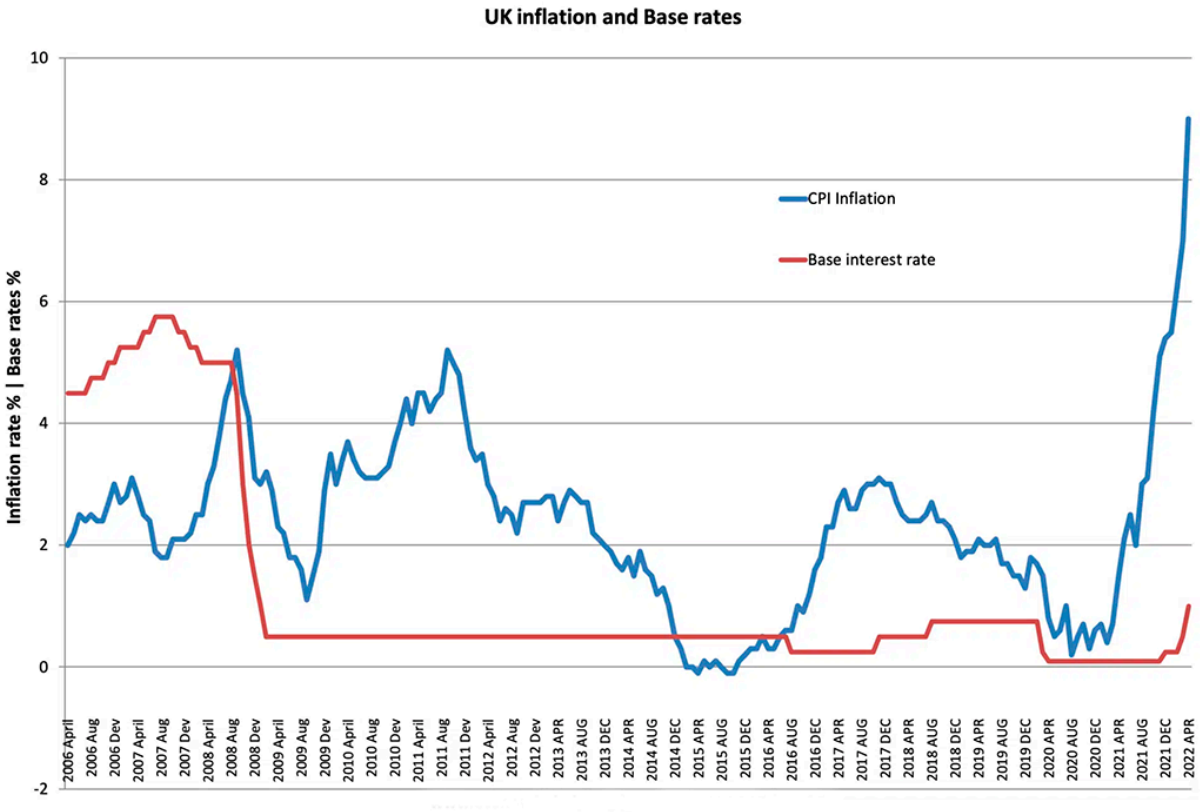
We could think that perhaps the income of London citizens has increased in a similar proportion to the increase in housing prices, but as we can see below, this has not been the case:



But, what is the reason for this price escalation?. According to expert british economists, the following could be mentioned:

Relatively low-interest rates

Low-interest rates make buying a home more attractive than renting. Also, low interest rates mean that buying a home can earn a better rate of return than buying other forms of investment, such as stocks. An investor looks at the return on housing (rental income) versus the cost of buying a home (mortgage interest payments). Very low interest rates increase the attractiveness of buying a home as an investment. Since 1992 interest rates in the UK have fallen from 15% to 0.5%, making the cost of getting a mortgage much lower.



During a period of high inflation (as we see in 2022) it also makes it attractive to buy a house as during periods of high inflation, houses (a real asset) will hold their value better than cash in a bank. Basically, the era of very low interest rates has been a key factor in pushing up house prices

Constraints on House Building/supply

Because of the growing number of households and growing demand for housing, the British government estimate Uk needs to build 250.000 new houses a year, just to keep pace with a rising population. Because house building is at its lowest level since the Second World War.

There are many constraints on the building of houses:

- In the most popular areas, there is a shortage of supply. It is difficult to find new land around Greater London.
- Environmental cost. The British have a strong attachment to preserving “Greenbelt Land” and many areas are protected from further housing development.
- Not in my back yard. People are usually in favor of more homes being built, as long as they are not in their local area. Increasing supply of houses leads to more congestion, crowded amenities and loss of greenbelt land.
- Vested interests perhaps most importantly increased supply reduces the value of your existing home. Therefore, existing homeowners have a vested interest in keeping the supply as low as possible in their area.
- Lack of Social Housing. Since Mrs Thatcher encouraged the sale of council housing, the number of new social housing (an euphemism for council housing) has been very low.

The consequence of this growing demand compared to limited growth in supply, is that there is strong economic pressure on house prices.

Demand is growing

A very simple economic truth: if demand increases faster than supply then prices will rise. The London population continues to grow (9.541.000 a 1,22% increase from 2021). Also, the number of householders is growing at a faster rate than the population.

Strong demand for home ownership

In recent years, the % of first-time buyers has fallen. The number of people able to buy a house has fallen, due to the decline in affordability. However, there is a strong cultural and economic desire to own your property. Increasingly common is for parents to help their children to buy a house, with a deposit or even putting the mortgage in their name. This has enabled first-time buyers to overcome the impossible income multiples and buy despite the expensive prices.

Speculation / Buy to let

Housing has increasingly been seen as a good investment. The returns on buying a house have consistently outperformed the stock market. In London, there has been a lot of demand from foreign nationals such as Russians and Arabs. Some argue this speculative increase in demand means the high house prices are unsustainable and are liable to fall.

Renting is also expensive

The alternative to buying a house is renting. But, the cost of renting has also risen faster than incomes. The increased price of renting reflects the fundamental imbalance in demand and supply. It is true that the price of housing is now rising faster than renting, but it still makes economic sense to buy rather than rent.

The Covid Effect

In March 2021 the Covid shutdown led to an unprecedented fall in GDP, with many workers losing income or their jobs. Despite GDP still being below the pre-crisis trend there has been no let up in the rise in UK house prices. Whilst those on furlow may struggle with rent, Covid has ironically made buying a house even more attractive.

Wealth inequality

There is substantial wealth inequality in the UK. With inheritance enabling some to get on the property ladder and afford seemingly ‘unaffordable’ prices. But, the inheritance effects makes it even harder for those who do not benefit from their parents.

After detailing the economic & demographic reasons that have led London to be one of the cities with the most expensive housing in Europe, we are going to try to develop a price prediction model that allows us to determine their evolution based on endogenous factors, that is, those specific to the attributes of a home.

Of course we will rely on the data provided officially by different authorities to determine the best predictors for a model whose response variable is the purchase price of a house for the districts of London. This is done by creating a model that describes the variability of the data with as few predictor variables as possible. And after the application of the model is finished, we hope that the following questions, beyond the factors that we have just explained, can be answered:

- is the age of the house relevant?
- how important is the surface of the house? and the rooms?
- is it only the surface or also the location that determines the price of a home?
- is the availability of a garage a relevant criterion when considering the purchase?
- having comforts such as central heat system really increases the value of a home? Let's see what the analysis we are going to carry out next, says.

3. Accesing our dataset - Data exploration

The first thing we will do is review the dataset that we have obtained from the official site of the City of London. We are going to check the data that it offers us and what it consists of.

The LonDat dataset have a total amount of 12.536 lines and 31 variables (*X*, *Longitud*, *Latitud*, *PrCompra*, *CXX1*, *CXX2*, *CXX3*, *CXX4*, *CXX5*, *CTCH*, *CTA*, *TPA*, *GSen*, *GDob*, *AlPro*, *Acond*, *DoBa*, *HabDos*, *HabTres*, *HabCuat*, *HabCin*, *CXXI*, *MCC*, *SinCo*, *Cper*, *ProfC*, *ProfNC*, *ResJub*, *DesVen*, *TasDes*, *DenPob*). The description of the variables is as follows:

Variable	Description
Longitud	Longitude in m
Latitud	Latitude in m
PrCompra	Purchase price in GBP
CXX1	Built between 1931 and 1960
CXX2	Built between 1961 and 1970
CXX3	Built between 1971 and 1970
CXX4	Built between 1981 and 1990
CXX5	Built between 1991 and 2000
CTCH	Detached property
CTA	Semi−detached property
TPA	Flat or apartment
GSen	Single Garage
GDob	Double Garage
AlPro	Leasehold/Freehold indicator
Acond	Central heating
DoBa	Two or more bathrooms
HabDos	Two bedrooms
HabTres	Three bedrooms
HabCuat	Four bedrooms
HabCin	Five bedrooms
CXXI	New property (2000s and advance)
MCC	Floor area in square metres
SinCo	Proportion of households without a car
Cper	Cars per person in neighbourhood
ProfC	Proportion of households with professional head
ProfNC	Proportion of households with unskilled head
ResJub	Proportion of residents retired
DesVen	Rate of unemployed who cannot buy a house
TasDes	Unemployed workers
DenPob	Local population density

The variables can be analysed as:

- *Geographical data* – Latitude, Longitude
- *Numerical data* – DoBa, HabCin, HabCuat, HabTres, HabDos, CXX3, CXX4, CXX5, CXX2, CXX2, CXXI, Acond, MCC, GDob, GSen, CTCH, CTA, TPA, AlPro
- *Neighbourhood* - Cper, ProfC, ProfNC, ResJub, DesVen, TasDes, DenPob

The above table shows complete list of variables organised into several groups, and for the most part representing the levels in a categorical variable expanded into dummy (0/1) variables.

- *Age* - Represent the time period in which the property was constructed. The omitted category is built before 1914.
- *Tenure* - Represents the type of building. The omitted category is bungalow.
- *Garage* - Omitted category is NO Garage.

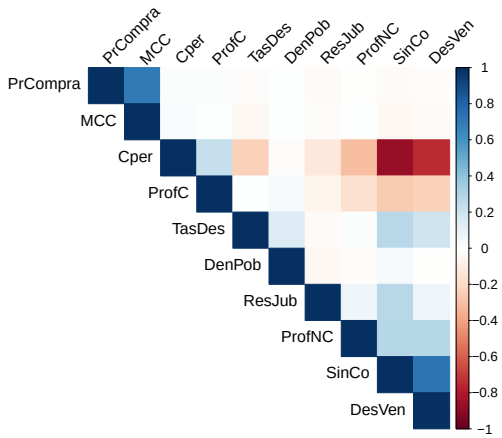
4. Correlation analysis

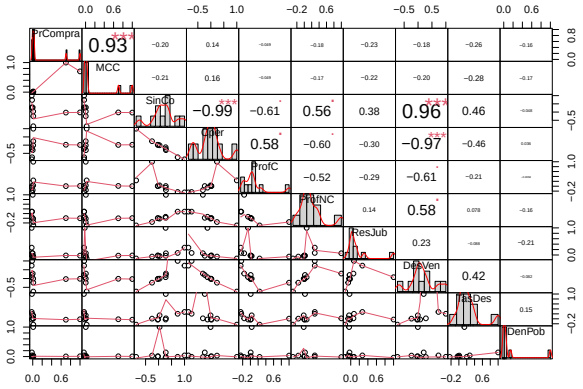
Correlation analysis is statistical method that is used to discover if there is a relationship between two variables/datasets, and how strong that relationship maybe. In terms of data analysis, this means that correlation analysis is used to analyse quantitative data gathered from research methods to identify whether there is any significant connections, patterns, or trends between the two.

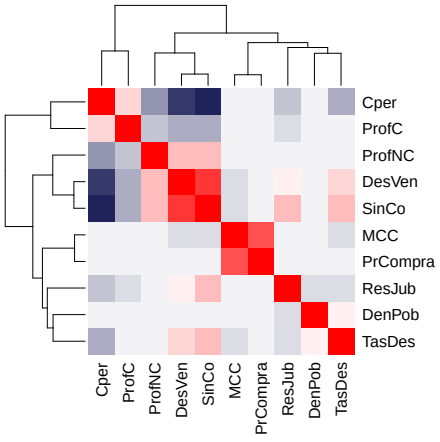
Essentially, correlation analysis is used for spotting patterns within datasets: a positive correlation result means that both variables increase in relation to each other, while a negative correlation means that as one variable decreases, the other increases.

In our case, the relationship between the variables/predictors help us to prevent any high influence predictor to deviate the results if it is not significant. Best way is to minimize the *collinearity* (linear relationship between two explanatory variables) between the variables to reduce the influence. In below graph, darker shades of blue and red conveys the high positive and negative association:

```
##          PrCompra   MCC SinCo  Cper ProfC ProfNC ResJub DesVen TasDes DenPob
## PrCompra     1.00  0.70 -0.02  0.01  0.01   0.00  -0.02  -0.01  -0.02   0.01
## MCC          0.70  1.00 -0.03  0.02  0.00   0.01  -0.02  -0.03  -0.04   0.00
## SinCo       -0.02 -0.03  1.00 -0.86 -0.25   0.28   0.27   0.73   0.28   0.04
## Cper         0.01  0.02 -0.86  1.00  0.24  -0.31  -0.13  -0.74  -0.23  -0.01
## ProfC        0.01  0.00 -0.25  0.24  1.00  -0.16  -0.06  -0.24   0.01   0.03
## ProfNC       0.00  0.01  0.28 -0.31 -0.16   1.00   0.06   0.28   0.02  -0.02
```







From previous figures we can conclude:

- *PrCompra (Purchase Price)* and *MCC (Floor area in square meters)* are positively correlated with value of 0.7014.
- *DesVen (Rate of unemployed who cannot buy a house)* and *SinCo (Proportion of houses without a car)* are positively correlated with value of 0.734.
- *Cper (Cars per person in neighborhood)* and *SinCo (Proportion of houses without a car)* are negatively correlated with value of -0.863.

- *DesVen (Rate of unemployed who cannot buy a house)* and *Cper (Cars per person in neighborhood)* are negatively correlated with value of -0.7400.

Now, we know the correlation between different predictors we have to bear in mind for our project, without neglecting the rest that have a slight relationship.

5. Data wrangling

As mentioned at Point *3.* we have noticed that there are four columns that had dummy expanded categorical variables for which few data were missing.

Since we have detected that there are dummy variables that represent a single entity, we decided to convert these dummy variables into factor variables:

New Variable	From Variables
Age	CXX1
	CXX2
	CXX3
	CXX4
	CXX5
	SXX – missing values
Type	CTCH
	CTA
	TPA
	Others – missing values
Garage	GSen
	Gdob
	Aparcam – missing values
Bedrooms	HabDos
	HabTres
	HabCuat
	HabCin
	HabUna – missing values

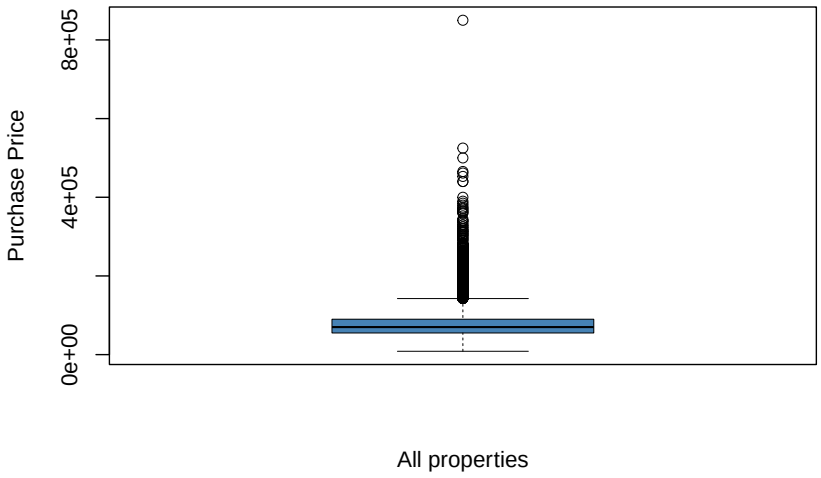
After the review and once converted dummies to factors, we get the final columns:

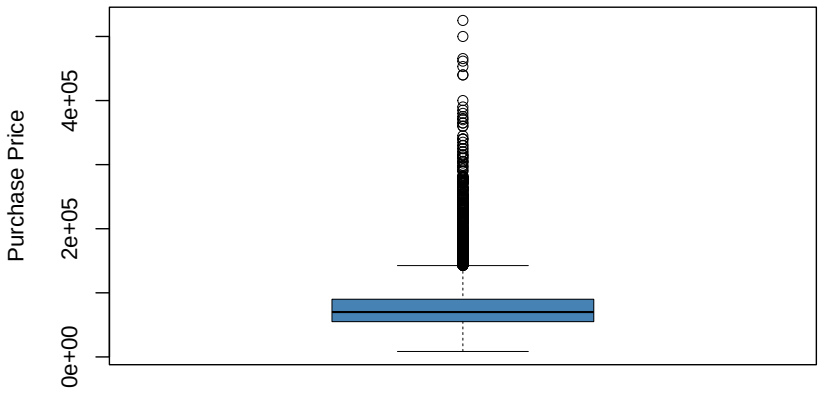
```
##   Longitud Latitud PrCompra AlPro Acond DoBa CXXI       MCC   ProfC  Age Type
## 1   545500  173000    85000   yes   yes   no   no  76.16146  0.0000 CXX3 CTCH
## 2   525000  177800    71000   yes   yes   no   no  98.45262  6.2500 CXX5 CTCH
## 3   531100  183400    60000   yes   yes  yes   no 124.73761  0.0000  SXX  CTA
## 4   538500  169400    64000   yes   yes   no  yes 127.00000  0.0000 CXX5 CTCH
## 5   534000  168400   260000   yes   yes  yes   no 190.40366  9.0909 CXX5 CTCH
## 6   528700  168800    48500   yes   yes   no   no  87.00000 16.6667  SXX  TPA
##    Garage Bedrooms
## 1    GSen  HabTres
## 2    GSen  HabTres
## 3 Aparcam  HabCuat
## 4    GSen  HabTres
## 5    GDob  HabCuat
## 6 Aparcam  HabTres
```

Outlier Analysis & House Prices

Outlier analysis involves identifying abnormal observations in a dataset. It helps to remove erroneous or inaccurate observations which might otherwise skew conclusions.

In the other had, Normalization is the process of ensuring that all of the data points in a dataset are formatted in the same way, so that they can be manipulated equally. Without normalization, it may be impossible to sort, graph, or otherwise assess datasets. Boxplot is a great way to detect outliers and inspect the variables by creating visual for the given variables:

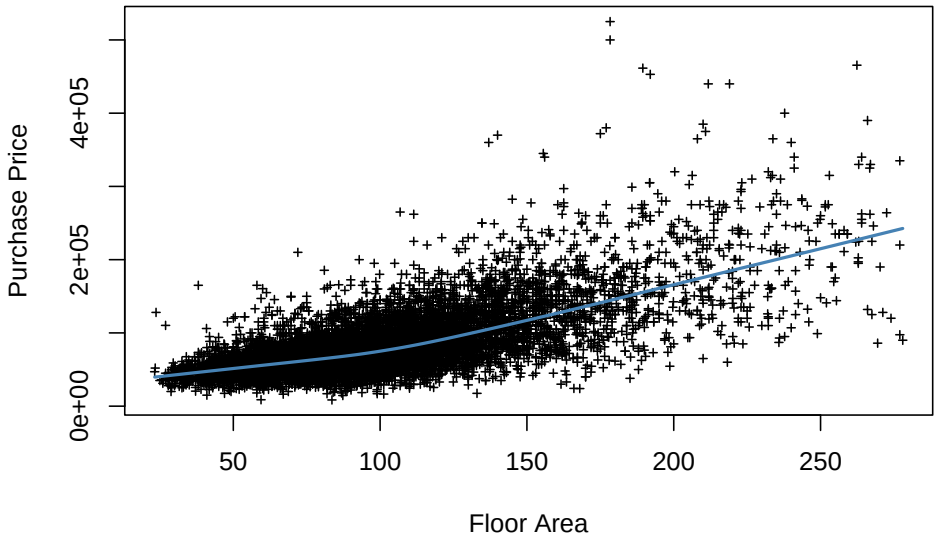




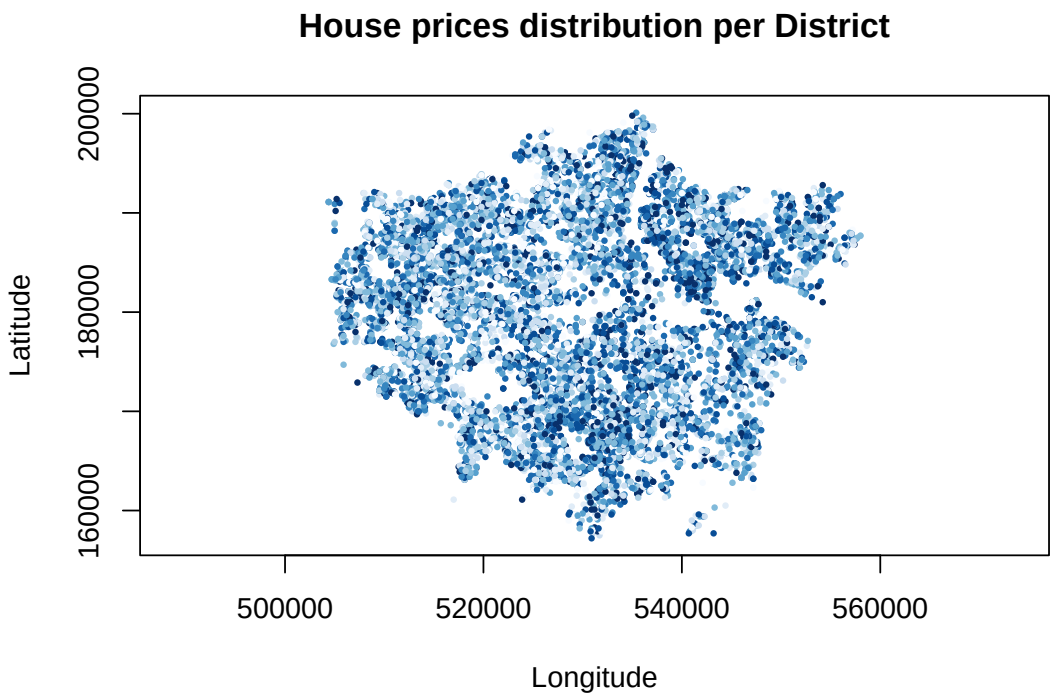
Properties under 600k

In the first boxplot, we notice a very large number of points outside the quantile regions. As we can see, there are single values above 600k GBP which could be an error. Thus, we decide to remove only the data points above 600k GBP. In the resulting boxplot, we can see a normalized result.

By examining the price of the house compared to the floor area, it can be seen that as the size of the house increases, so does the price:



However, there are still many outliers as evidenced by the box plot. We consider other variables may be taken into account for the discrepancy, as house prices do not depend solely on their size. Geographical for example as properties in one district may not cost the same as in other districts, even if they have the same floor square meters:



The house price map of London districts, shows how these are distributed. By applying a model that relates this price to the east and north extremes, we can see that prices increase moving north and west, while prices decrease moving south and east. However, and despite this first consideration, we will proceed to add other predictor variables to determine if this is the only relationship to consider in terms of housing prices:

```
## [1] 302279.5
```

```
## [1] 302256.2
```

```
##
## Call:
## lm(formula = PrCompra ~ Long + Lat, data = Project)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -72993 -25021 -10060   9656 769670
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 165718.67   17913.60   9.251  < 2e-16 ***
## Long          -134.20      30.84  -4.351 1.37e-05 ***
## Lat            -82.51      41.01  -2.012   0.0442 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41660 on 12533 degrees of freedom
## Multiple R-squared:  0.001824,   Adjusted R-squared:  0.001664
## F-statistic: 11.45 on 2 and 12533 DF,  p-value: 1.077e-05
```

```
##
## Call:
## lm(formula = PrCompra ~ Long + Lat + I(Long^2) + I(Lat^2) + I(Long *
##     Lat), data = Project)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -74048 -24883  -9900   9821 768598
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.300e+06  8.862e+05  -3.724 0.000197 ***
## Long           1.273e+04  2.832e+03   4.495 7.04e-06 ***
## Lat            5.999e+02  2.956e+03   0.203 0.839190
## I(Long^2)     -1.118e+01  2.591e+00  -4.317 1.60e-05 ***
## I(Lat^2)       6.694e+00  4.782e+00   1.400 0.161572
## I(Long * Lat) -5.743e+00  4.383e+00  -1.310 0.190152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41620 on 12530 degrees of freedom
## Multiple R-squared:  0.004151,   Adjusted R-squared:  0.003754
## F-statistic: 10.45 on 5 and 12530 DF,  p-value: 5.075e-10

## Start:  AIC=266678.6
## PrCompra ~ Long + Lat + I(Long^2) + I(Lat^2) + I(Long * Lat)
##
##                 Df  Sum of Sq        RSS    AIC
## - Lat            1 7.1327e+07 2.1702e+13 266677
## - I(Long * Lat)  1 2.9732e+09 2.1705e+13 266678
## - I(Lat^2)       1 3.3942e+09 2.1705e+13 266679
## <none>                        2.1702e+13 266679
## - I(Long^2)      1 3.2271e+10 2.1734e+13 266695
## - Long           1 3.4988e+10 2.1737e+13 266697
##
## Step:  AIC=266676.6
## PrCompra ~ Long + I(Long^2) + I(Lat^2) + I(Long * Lat)
##
##                 Df  Sum of Sq        RSS    AIC
## <none>                        2.1702e+13 266677
## - I(Lat^2)       1 6.5347e+09 2.1709e+13 266678
## - I(Long * Lat)  1 6.7635e+09 2.1709e+13 266679
## - I(Long^2)      1 3.2570e+10 2.1735e+13 266693
## - Long           1 3.9453e+10 2.1741e+13 266697

##
## Call:
## lm(formula = PrCompra ~ Long + I(Long^2) + I(Lat^2) + I(Long *
##     Lat), data = Project)
##
## Coefficients:
##   (Intercept)           Long      I(Long^2)       I(Lat^2)  I(Long * Lat)
##    -3.188e+06      1.251e+04     -1.110e+01      7.295e+00     -5.018e+00
```

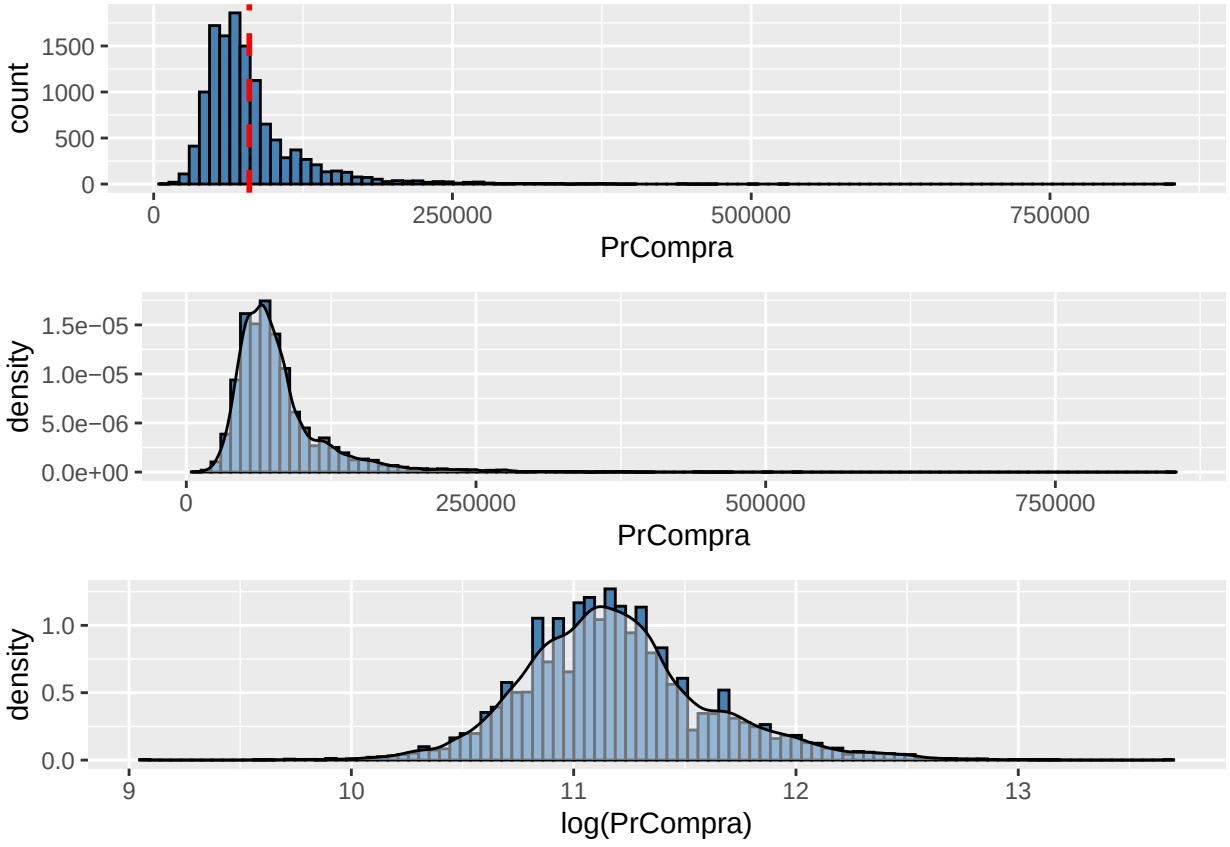
After applying the relationship model between housing price and geographic location, we can confirm what was seen previously and affirm that, as we see, by having a lower AIC, prices increase when moving towards the North and West, while they decrease when move south and east.

5. Data distribution

Purchase Price distribution

The main variable we´d like to analyze is *PrCompra* (Purchase Price), as is the one we want to predict. We´re going to proceed as follows:

- First we will construct a histogram to visualize the distribution of *PrCompra* and find the mean value (dashed line).
- Second, we will apply a density plot allowing us to have smoother distributions by smoothing out the noise. The peaks the density plot, will help us display where values are concentrated over the interval.
- Third, we´re going to treat the variable with a logarithmic transformation, as data is clearly right skewed and it will allow us to give it a more normal pattern.



As we can see, the third plot for *log(PrCompra)*, clearly shows that *PrCompra* displays normality.

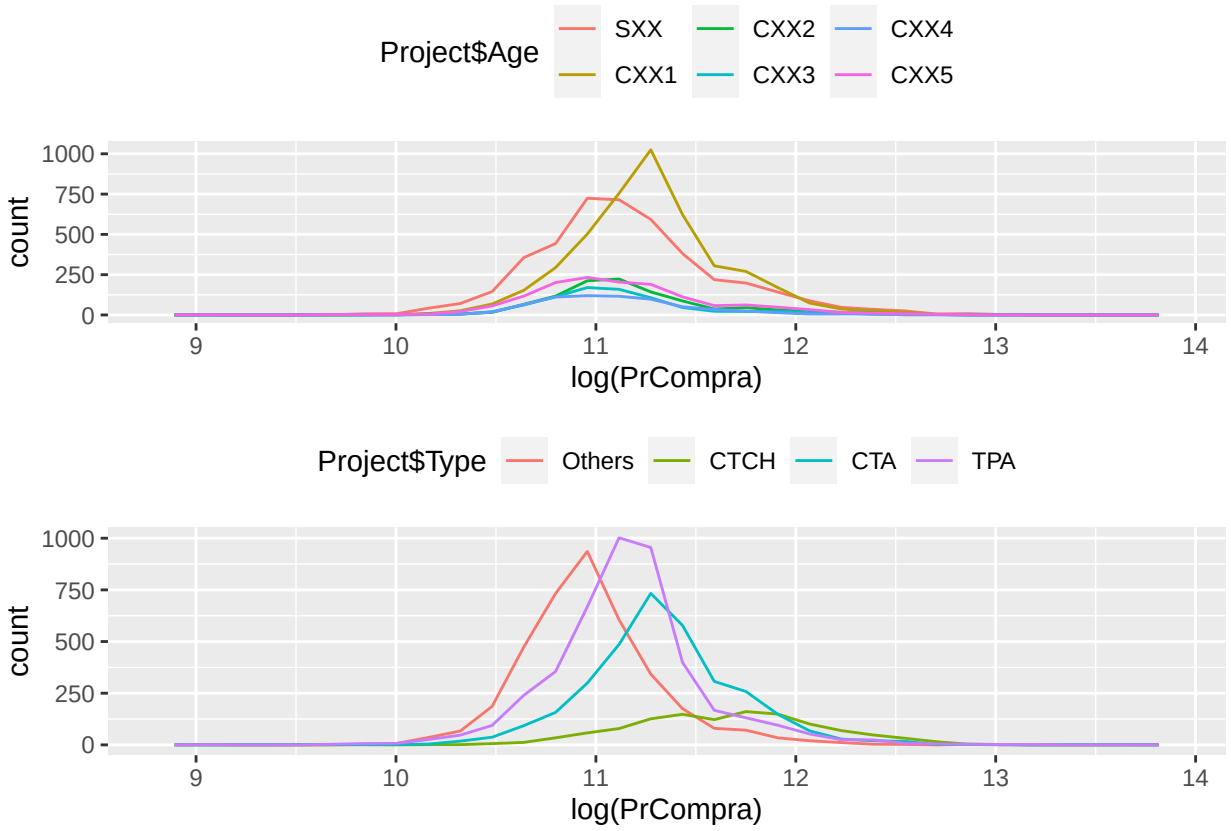
Predictors distribution

The rest of the variables in the dataset have a categorical character (*has or does not have*) such as *Acond* that defines if the house has central heating or not, depicted as a dummy variable. In the other hand, we have multiple dummy variables representing a single category like *Age*, *Type* or *Bedrooms* that have been merged into factor columns. In

order to be able to obtain clear information from these predictors versus *PrCompra* (Purchase Price), we are going to treat them with a logarithmic transformation again:

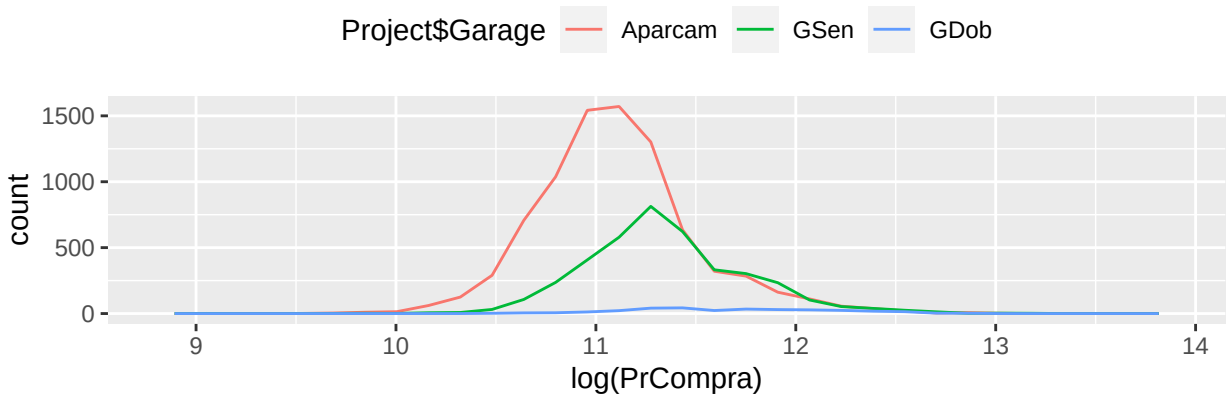
First comparison

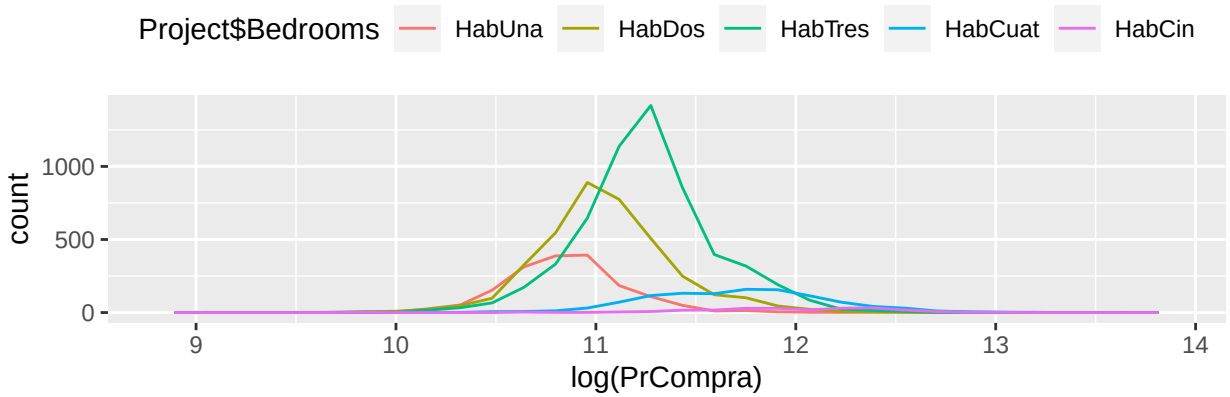
- *PrCompra (Purchase Price) vs Age*: this plot depicts that properties 'Built between 1931 and 1960'(CXX1) have the highest purchase ratio among all the categories, followed by 'Built previously 1930'(SXX). The rest of the variables are shown very far from these two.
- *PrCompra (Purchase Price) vs Type*: this plot shows that people are willing to purchase flat or apartments most preferably, followed by other type of properties and at the end, semi-detached.



Second comparison

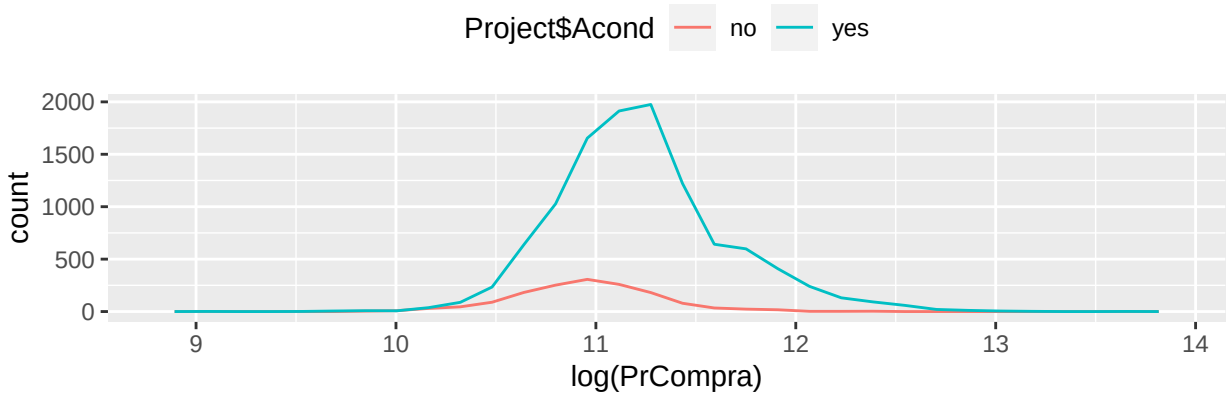
- *PrCompra (Purchase Price) vs Garage*: The line plot shows that properties with garage are in the majority, followed by those with a single garage.
- *PrCompra (Purchase Price) vs Bedrooms*: From this graph we can see that the vast majority of buyers decide on those homes that have three bedrooms. In decreasing order, they opt for those with two and one bedrooms. We also see that homes with four and five bedrooms are the least sold.

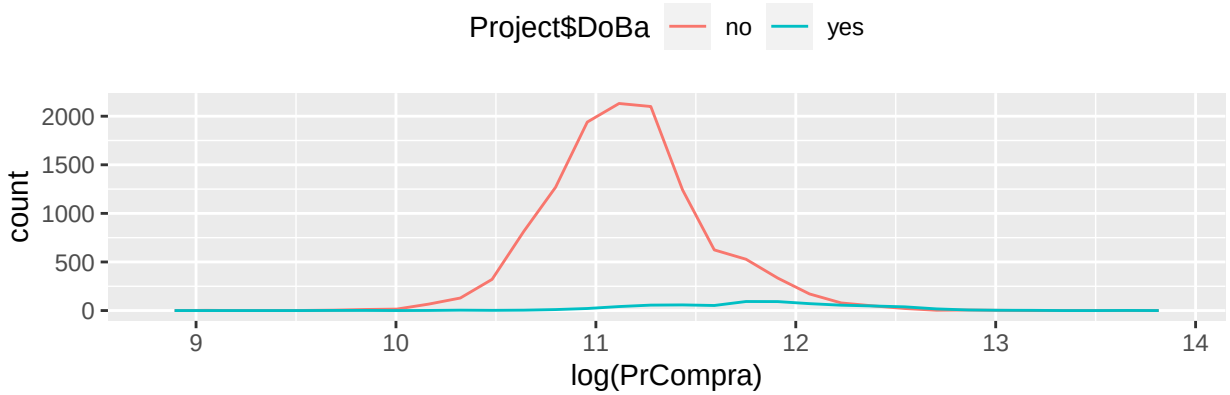




Third comparison

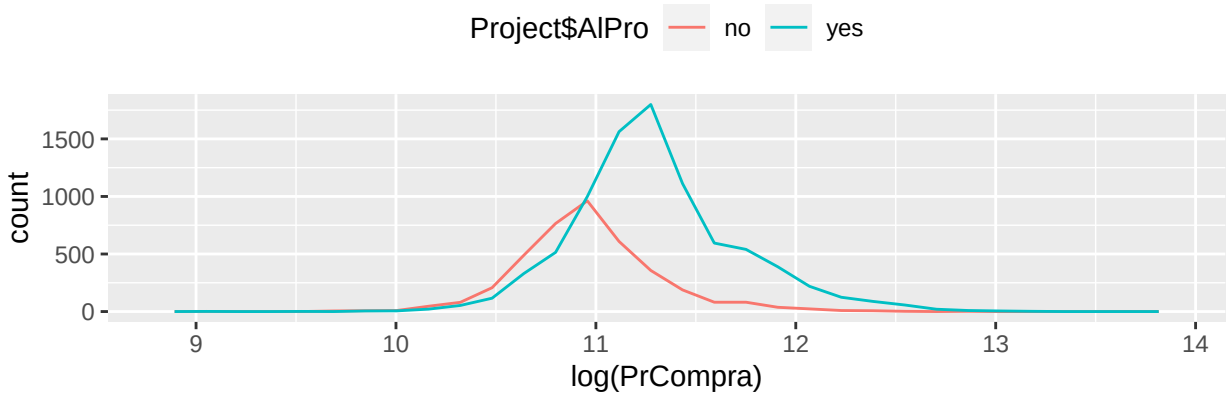
- *PrCompra (Purchase Price) vs Acond*: In this plot, we can see that having an extra comfort, such as central heating, is an added value to a home, since they are the most selected by the owners.
- *PrCompra (Purchase Price) vs DoBa (Two or more bathrooms)*: This graph tells us that homes with two bathrooms or more are in the minority as opposed to those with only one bathroom.

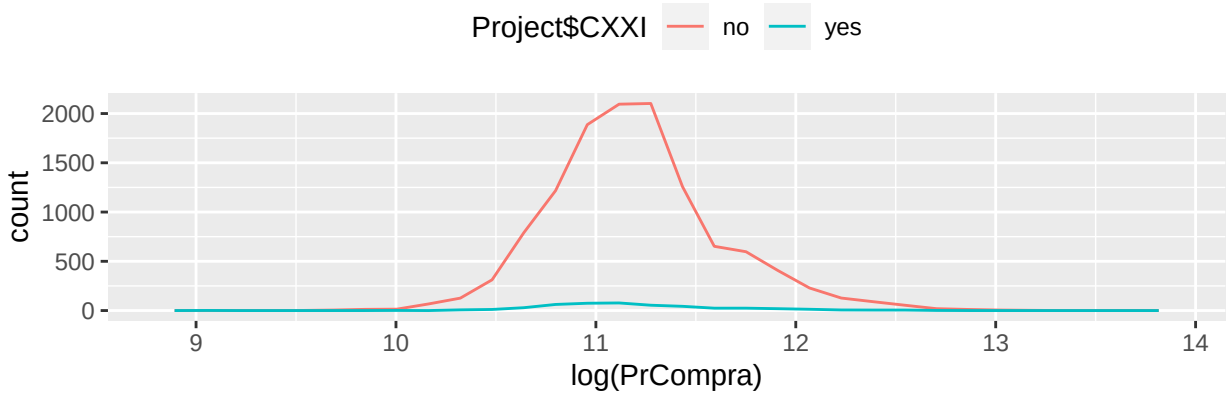




Fourth comparison

- *PrCompra (Purchase Price) vs AlPro (Leasehold or Freehold)*: In this graph we can see that the properties under lease are the majority if we compare them with those that are owned.
- *PrCompra (Purchase Price) vs CXXI (New property-S.XXI)*: In this case we can see that the vast majority of homes are new compared to those that were built before the 21st century.





6. Modelling

Predictors relevance - AIC

The fact of being able to determine the correlation that exists between the predictors and what we obtain from this interaction leads us to apply an estimation to the data model. If we take into account that the response variable is continuous (numerical variables that have an infinite number of values between any two values), we can affirm that the regression favors the modeling.

We have decided to apply the regression on the *PrComp* variable linked to each of the other predictors. In order to determine the most appropriate estimate, we will apply the AIC. AIC is an estimator (a single number score) that can be used to determine which of multiple models is most likely to be the best model for a given dataset. It estimates models relatively, meaning that AIC scores are only useful in comparison with other AIC scores for the same dataset. A lower AIC score is better.

Through the AIC estimator, we will be able to determine the model closest to what we are trying to determine.

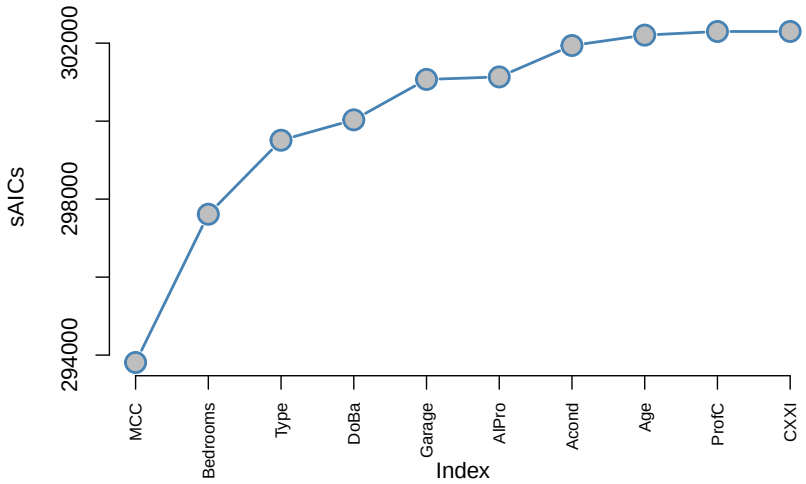
```
##  [1] 301133.9 301938.9 300032.7 302299.4 293810.5 302298.6 302205.9 299509.6
##  [9] 301069.6 297612.0
```

```
## [1] 293810.5
```

```
## [1] "MCC"
```

```
##
## Call:
## lm(formula = formula(paste0("PrCompra~", Vars[i])), data = Project)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -132969  -15467   -1430   11769  712918
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4020.849    739.104    5.44 5.42e-08 ***
## MCC          787.584      7.149  110.17  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29720 on 12534 degrees of freedom
## Multiple R-squared:  0.492,  Adjusted R-squared:  0.4919
## F-statistic: 1.214e+04 on 1 and 12534 DF,  p-value: < 2.2e-16
```

```
##      MCC Bedrooms     Type     DoBa   Garage    AlPro    Acond      Age
## 293810.5 297612.0 299509.6 300032.7 301069.6 301133.9 301938.9 302205.9
##    ProfC     CXXI
## 302298.6 302299.4
```




```
## Bedrooms 3801.50016469887
## Type 1897.6381998327
## DoBa 523.099352479971
## Garage 1036.86329427169
## AlPro 64.3165624550893
## Acond 804.993152776558
## Age 267.011313270836
## ProfC 92.7507540592924
## CXXI 0.738774296594784
```

When applying the AIC, we can see that the predictor *MCC (Floor Area)* is the most relevant, followed by *Bedrooms*. Other predictors add a slight variation to the model. *CXXI (New Property)*, *ProfC (Proportion of houses with professional qualification)* and *Age* are the least significant variables.

Linear Regression

All the analysis we have carried out so far allows us to conclude that there is a clear relationship between the predictors and the *PrCompra* variable. Among all the available regressions (Linear, Logistic, Polynomial…) we are going to apply the Linear model and fit it with all the predictors that exist in our dataset, in order to be able to explain the predictions and variations in the response variable, which could be attributed to changes in the variables.

We are not going to consider the variables Latitude and Longitude, since these are coordinates. However, we do not discard them, since they will be used later in the spatial analysis of the dataset.

```
##
## Call:
## lm(formula = log(PrCompra) ~ AlPro + Acond + DoBa + CXXI + MCC +
##     ProfC + Age + Type + Garage + Bedrooms, data = Project)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.14025 -0.15618  0.01117  0.16245  1.86198
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     10.3482881  0.0125495 824.597  < 2e-16 ***
## AlProyes         0.1271132  0.0142876   8.897  < 2e-16 ***
## Acondyes         0.1716804  0.0079722  21.535  < 2e-16 ***
## DoBayes          0.1566336  0.0127027  12.331  < 2e-16 ***
## CXXIyes          0.0429826  0.0163228   2.633 0.008467 **
## MCC              0.0063677  0.0001201  53.003  < 2e-16 ***
## ProfC            0.0003561  0.0002521   1.413 0.157805
## AgeCXX1          0.0476152  0.0069394   6.862 7.13e-12 ***
## AgeCXX2         -0.0342311  0.0103014  -3.323 0.000893 ***
## AgeCXX3         -0.0907449  0.0115122  -7.882 3.47e-15 ***
## AgeCXX4         -0.0814115  0.0123013  -6.618 3.79e-11 ***
## AgeCXX5         -0.0046106  0.0108430  -0.425 0.670690
## TypeCTCH         0.0032103  0.0175222   0.183 0.854635
## TypeCTA         -0.0915651  0.0152241  -6.014 1.86e-09 ***
## TypeTPA         -0.1655386  0.0147482 -11.224  < 2e-16 ***
## GarageGSen       0.0593897  0.0064923   9.148  < 2e-16 ***
## GarageGDob       0.0811809  0.0177092   4.584 4.60e-06 ***
## BedroomsHabDos   0.0380594  0.0091816   4.145 3.42e-05 ***
## BedroomsHabTres  0.0385969  0.0112841   3.420 0.000627 ***
## BedroomsHabCuat  0.0733782  0.0162889   4.505 6.70e-06 ***
## BedroomsHabCin   0.0276199  0.0264534   1.044 0.296461
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2869 on 12515 degrees of freedom
## Multiple R-squared:  0.5399, Adjusted R-squared:  0.5391
## F-statistic: 734.2 on 20 and 12515 DF,  p-value: < 2.2e-16
```

From the results obtained from the linear regression, we can observe the following:

- Intercept coefficient: 10,3488
- Parameters with significant p-values(<0,05) and positive co-efficients:
  - o CXXIyes
  - o MCC
  - o ProfC
  - o AgeCXX1
  - o TypeCTCH
  - o BedroomsHabDos
  - o BedroomsHabTres
  - o BedroomsHabCin
- *PrCompra* increases when the above predictors increase.
- Parameters with significant p-values(<0.05) and negative co-efficients are:
  - o AgeCXX2
  - o AgeCXX5
  - o TypeTPA
  - o TypeTypSemiD
  - o TypetypFlat
- *PrCompra* decreases when the above predictors increase.
- Parameters which are not significant p-values(>0.05) are:
  - o AlProyes
  - o Acondyes
  - o DoBayes o GarageGSen
  - o GarageGDob
  - o BedroomsHabCuat

The adjusted R-Squared is only 0.5391 (53%), with what we can consider a good result.

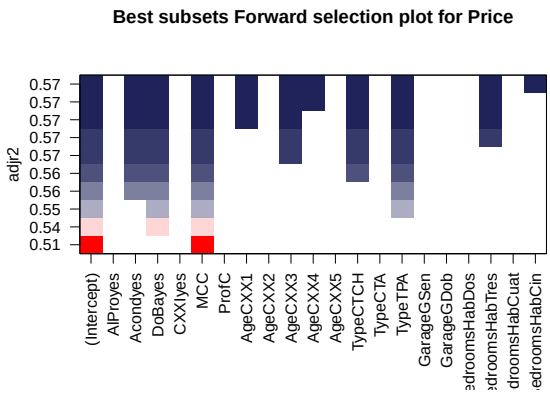
Best Subsets Regression

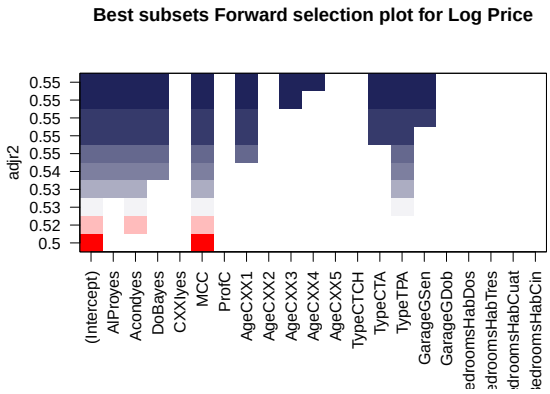
Best Subsets Regression is a model selection approach that consists of testing all possible combination of the predictor variables, and then selecting the best model according to statistical criteria.

First, we use a direct subset selection method for *PrCompra* and plot *adjr2* (adjusted coefficient of determination of a multiple linear regression model) for all variables in our previous Linear Regression Model.

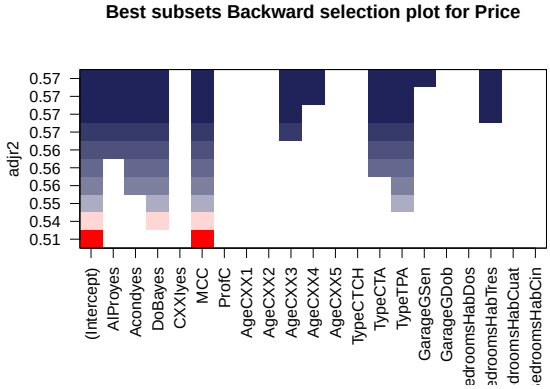
On the other hand, we use the selection plot best subsets backward selection plot for *PrCompra* and plot *adjr2* results for all variables. What we expect is that *MCC (Floor Area)* is the most significant variable of all.

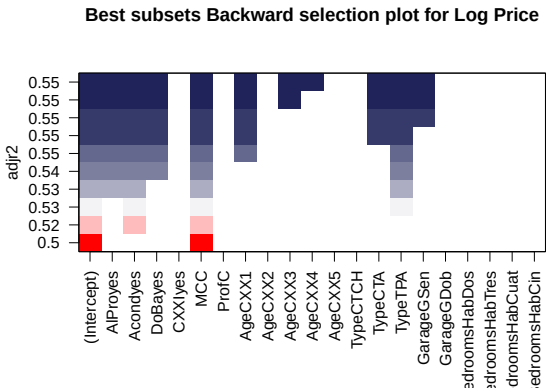
The best direct subset selection plot for *PrCompra* and *Log PrCompra* results for all variables are showing below:





In the other hand, we plot the backward selection of the best subsets for *PrCompra* and *Log PrCompra*:





What we can see in the plots for *PrCompra* is:

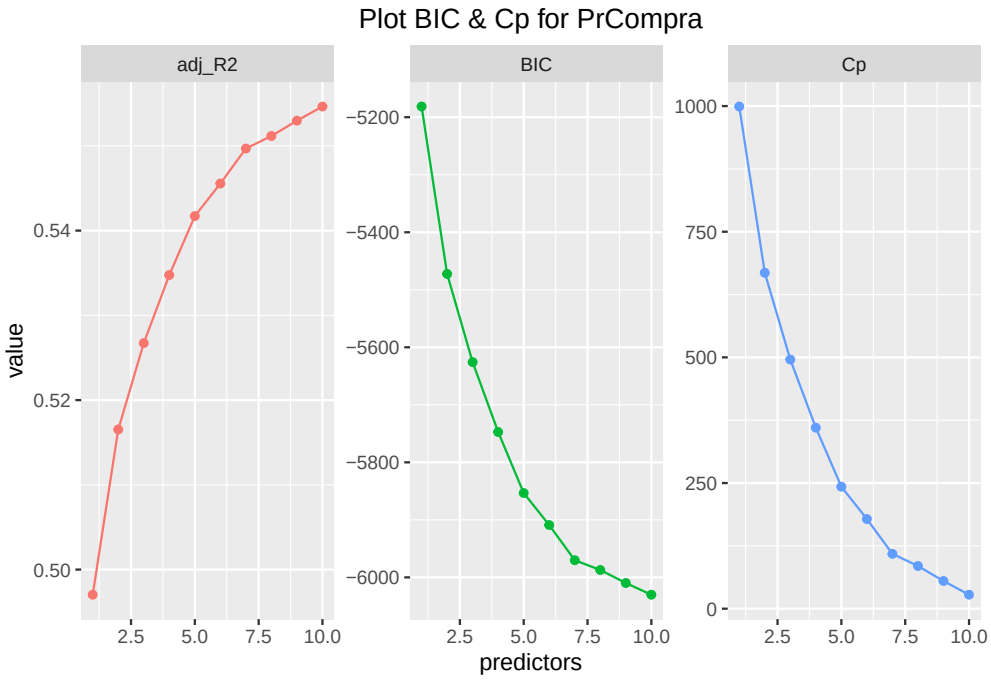
- MCC is the most significant variable.
- Acond is another relevant variable.

However, we can see certain differences in both approaches.

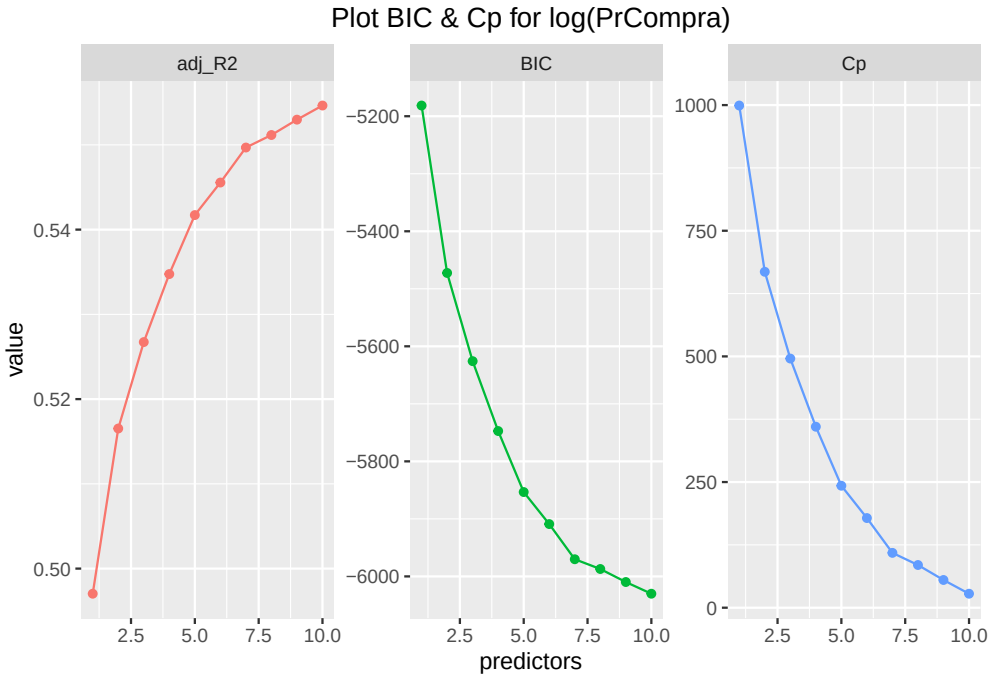
In the other hand, plots for *Log PrCompra* we can see that they provide us with similar significant variables Unlike the results obtained in Linear Regression model. From most to least relevant we got:

- MCC (Floor Area)
- Acond (Central Heat)
- TPA (Type Flat/Apartment)
- AlPro (Leasehold or Freehold indicator)
- DoBa (Two or more bathrooms)
- AgeCXX1 (Built between 1930 and 1960)
- CTA (Semi-detached property)
- GSen (Single Garage)
- AgeCXX3 (Built between 1971 and 1980)
- AgeCXX4 (Built between 1981 and 1990)

Following this approximation, we plot the *BIC (Bayesian Information Criteria)* and *Mallow´s Cp* plots for *PrCompra*. From this, we can identify that the value of *adjR2* is maximum at 10 and the values of *BIC* and *Cp* are minimum in 10, which implies that we can have up to 10 variables to create our model.



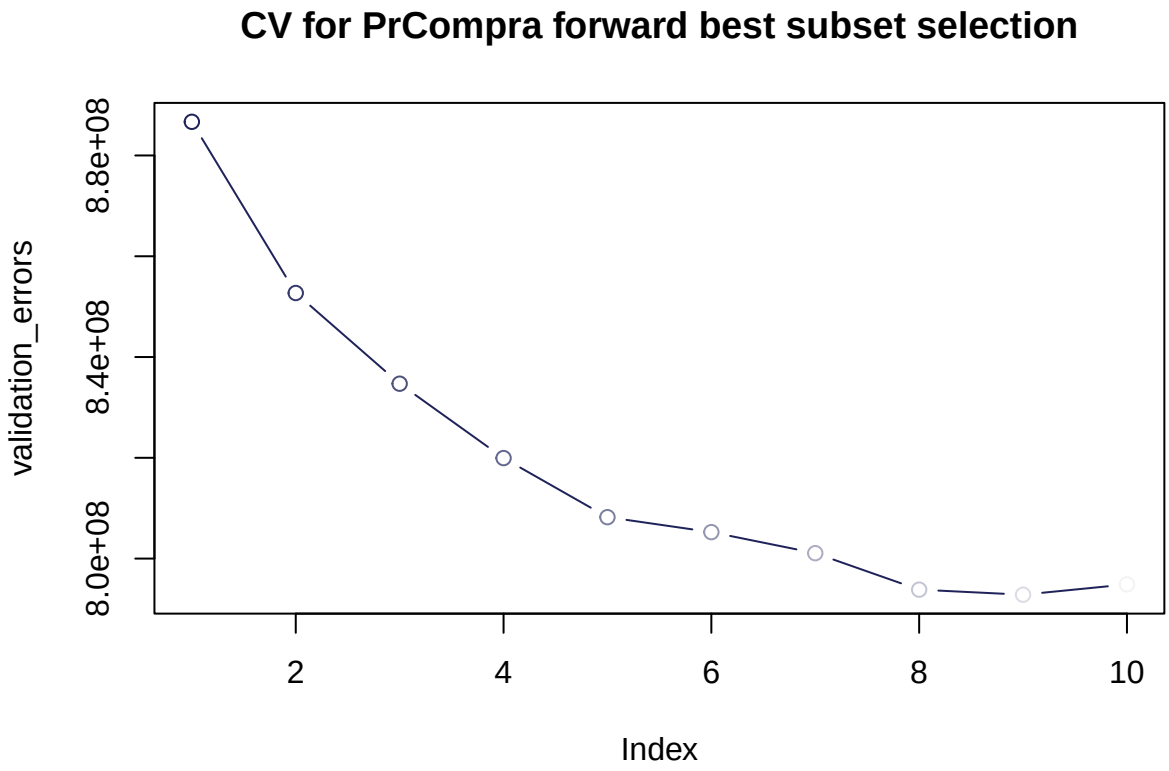
Plotting the *BIC* and *Cp* graphs for *Log PrCompra*, we get that the *adjR2* value is maximum at 10 and the *BIC* and *Cp* values are also minimum at 10, implying the same conclusion in the graphic for *PrCompra*. The only difference between the two is a slight change in curvature:

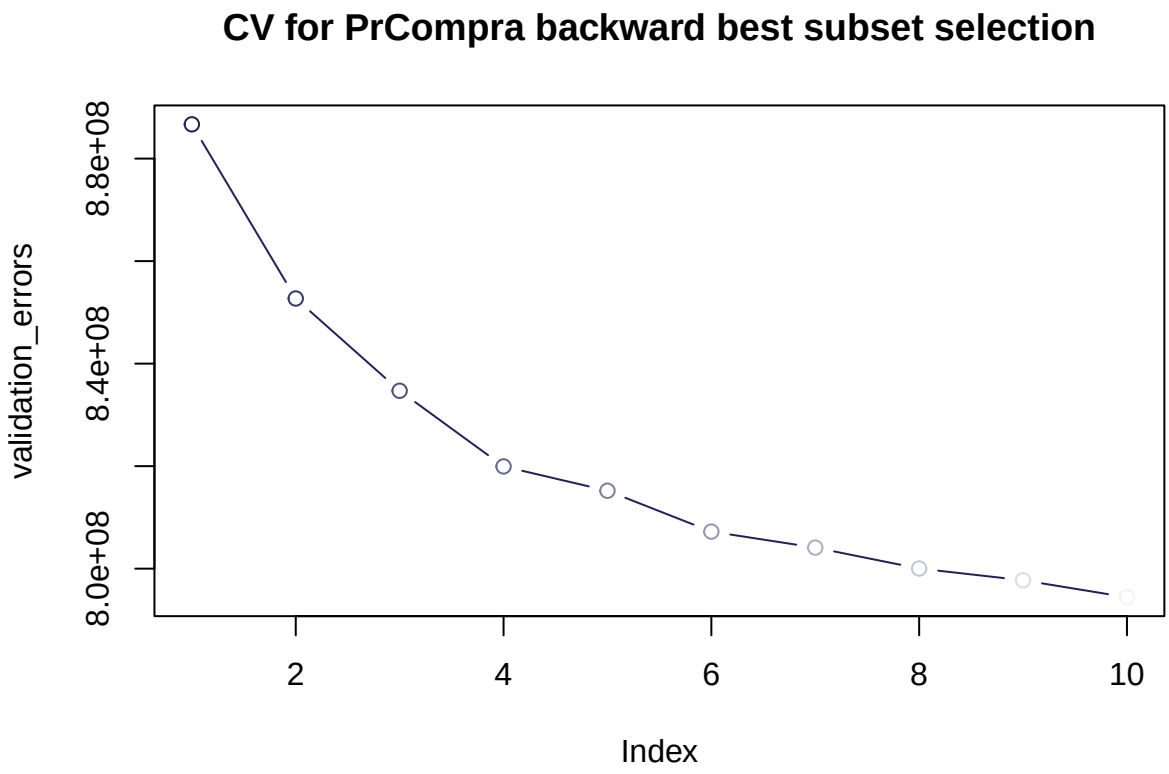


Cross Validation

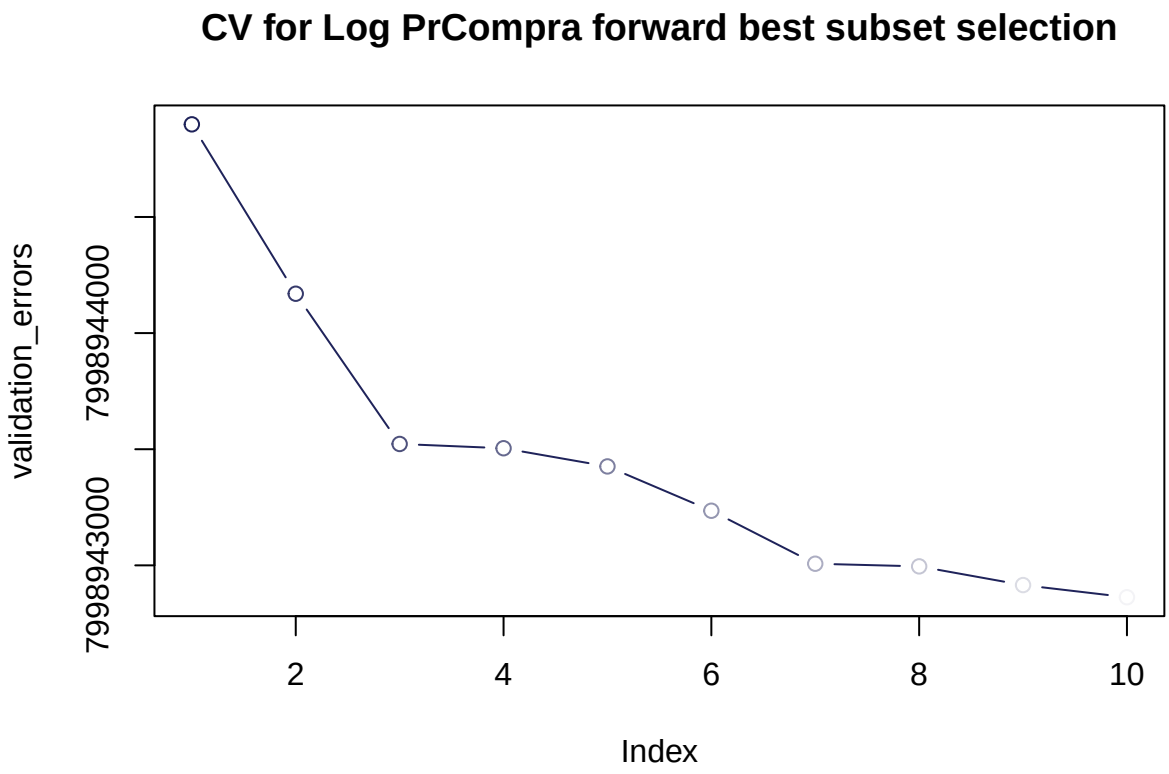
We are going to use the models previously created from the *training* model and apply them to the *test* model, verifying the errors resulting from the *Cross Validation*. We take into account that as we have previously determined in our

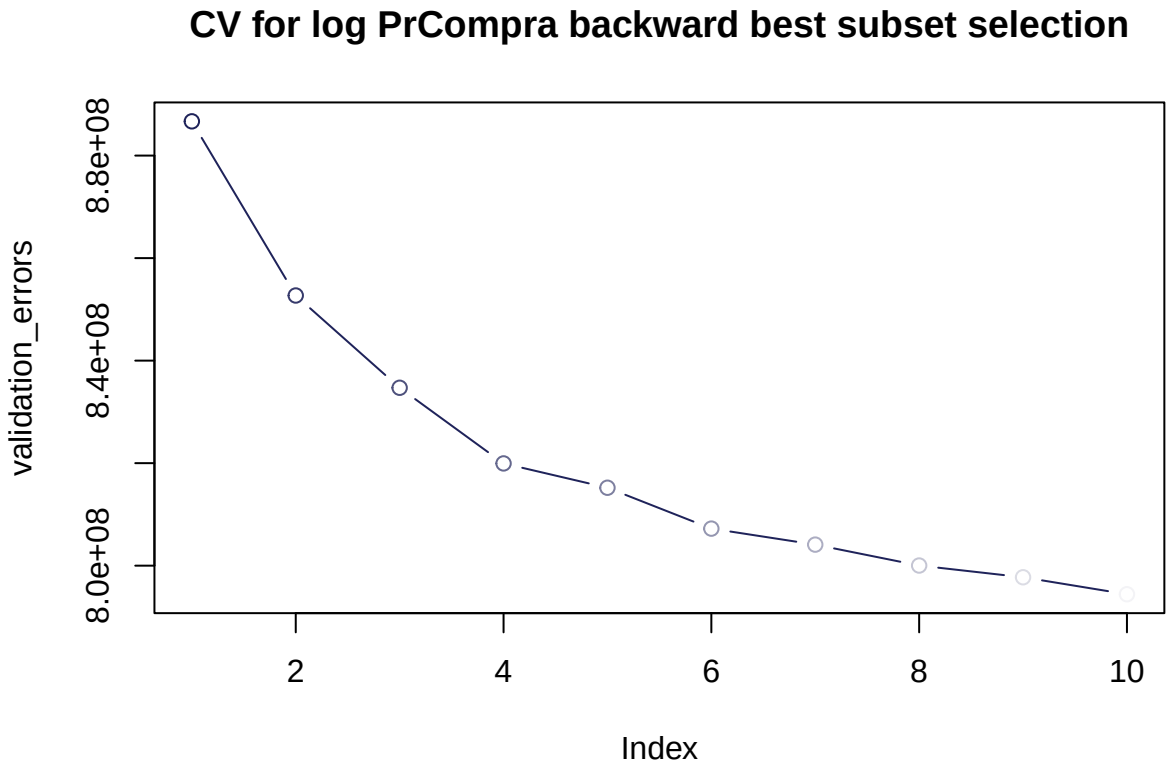
models, the selection of 10 variables is suggested to draw the graph.





If we now apply log, we will appreciate that when we plotting the cross validation errors the curvature varies from the two plots where log is not applied to *PrCompra*. Despite this, we can continue to consider that 10 is the number of variables that we can select to continue to be significant:


CV for Log PrCompra forward best subset selection
validation_errors
7998943000
7998944000
2
4
6
8
10
Index



From the information obtained so far, we can clearly say that MCC (Floor Area) has a determining role in predicting the final price of a home as opposed to the rest of the predictors. For our project, and after all approaches used, it is clear that the final model is definitely the following:

```
##
## Call:
## lm(formula = PrCompra ~ MCC + Bedrooms + Type + DoBa + Garage +
##     AlPro + Acond + Age + ProfC, data = Project)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -136573  -13454   -1379   10376  699087
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5970.77    1218.92   4.898 9.78e-07 ***
## MCC                683.85      11.67  58.604  < 2e-16 ***
## BedroomsHabDos   -3382.91     891.80  -3.793 0.000149 ***
## BedroomsHabTres  -7615.50    1096.02  -6.948 3.88e-12 ***
## BedroomsHabCuat  -2065.46    1582.05  -1.306 0.191727
## BedroomsHabCin    3125.56    2569.38   1.216 0.223832
## TypeCTCH          5292.46    1701.31   3.111 0.001870 **
## TypeCTA          -6990.13    1478.49  -4.728 2.29e-06 ***
## TypeTPA         -11873.38    1432.32  -8.290  < 2e-16 ***
## DoBayes          25140.48    1233.15  20.387  < 2e-16 ***
```

```
## GarageGSen        3723.20     630.50   5.905 3.61e-09 ***
## GarageGDob        9011.43    1719.98   5.239 1.64e-07 ***
## AlProyes          5950.71    1387.26   4.290 1.80e-05 ***
## Acondyes         11881.79     774.32  15.345  < 2e-16 ***
## AgeCXX1           3968.97     673.99   5.889 3.99e-09 ***
## AgeCXX2          -1249.28    1000.57  -1.249 0.211844
## AgeCXX3          -7490.11    1118.17  -6.699 2.20e-11 ***
## AgeCXX4          -6787.56    1194.81  -5.681 1.37e-08 ***
## AgeCXX5            861.39     922.20   0.934 0.350290
## ProfC               41.09      24.48   1.678 0.093307 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27860 on 12516 degrees of freedom
## Multiple R-squared:  0.5541, Adjusted R-squared:  0.5535
## F-statistic: 818.7 on 19 and 12516 DF,  p-value: < 2.2e-16
```

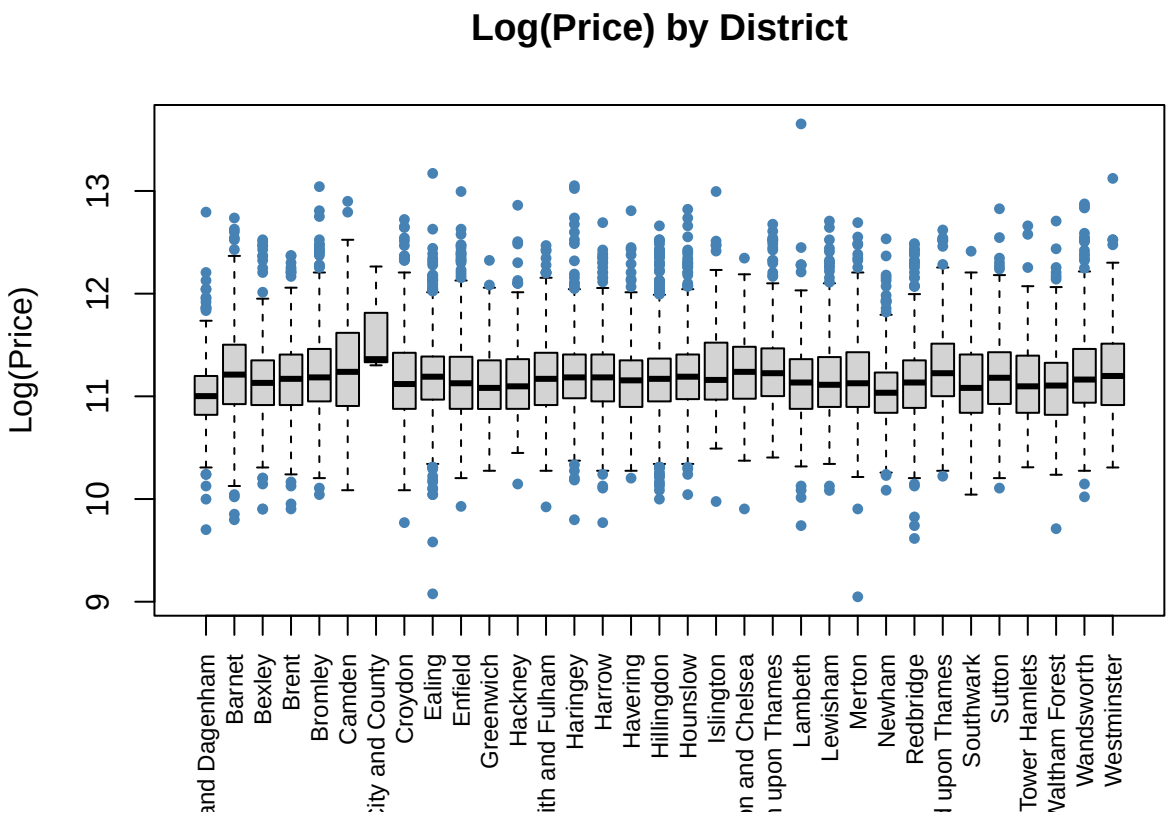
Geospatial analysis

Figure 1: London Boroughs (Districts)

Through geospatial analysis, we are going to apply a series of analysis methods that will allow us to establish the

relationship of the variables that we will establish with their geographical location. The variables chosen for this analysis are Municipality (District), Purchase Price, Floor Area, and Standardized Residuals.

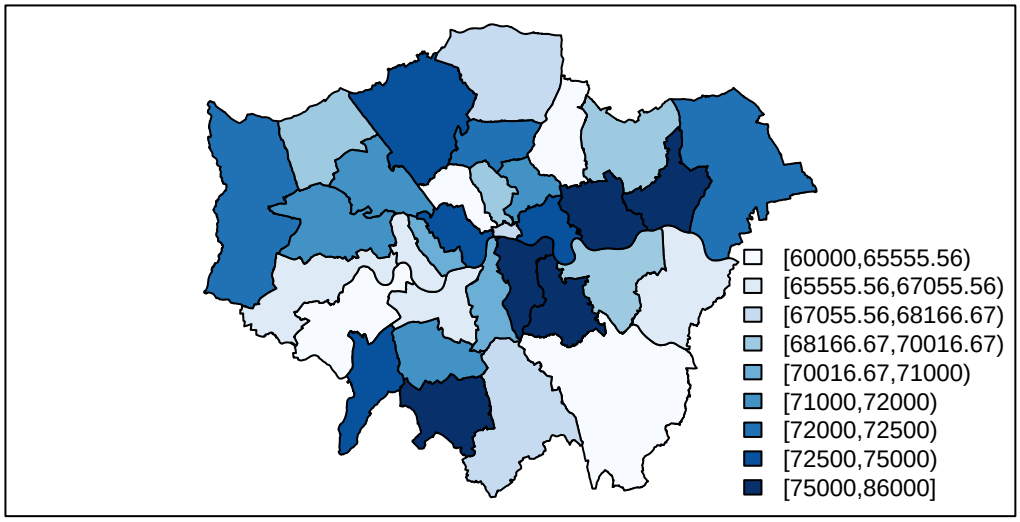
In order to understand if the Municipality (District) variable has an influence on the house price, following plot provide us relevant information:

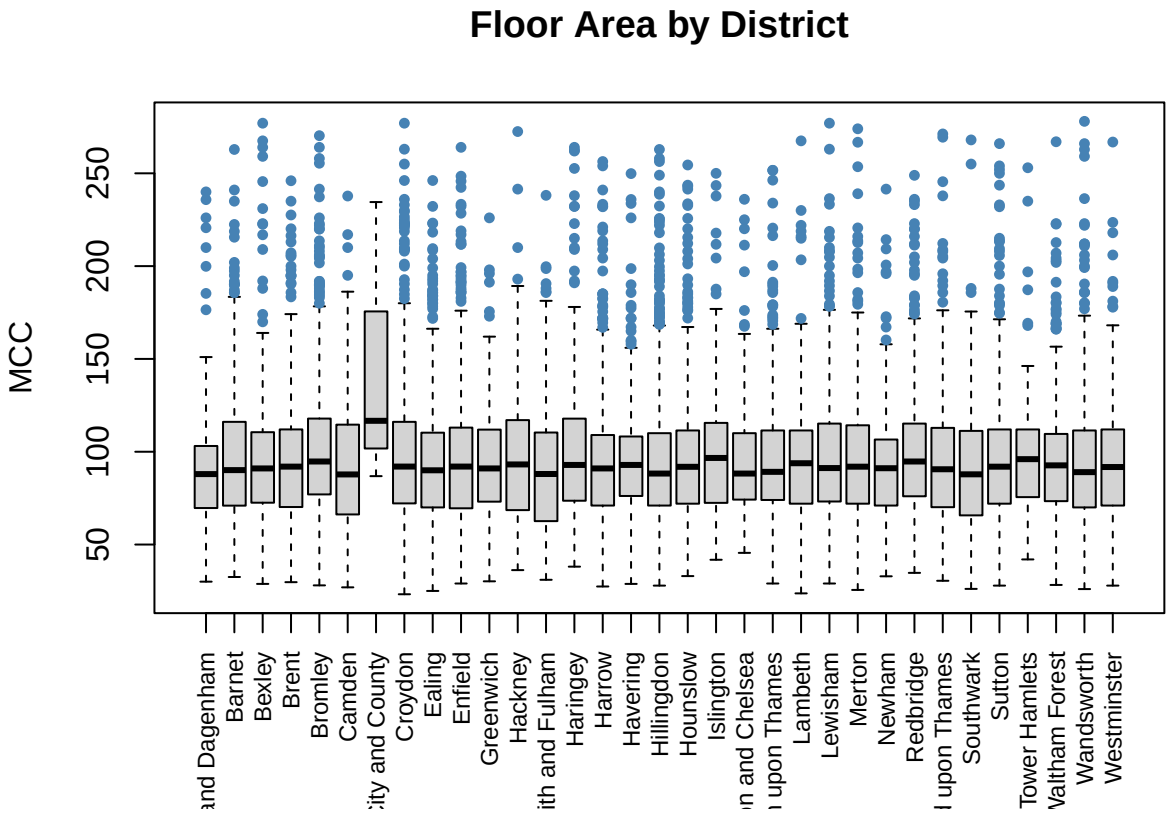


We can conclude that the median house price is roughly in the same range except for the City of London, where the median price is higher than elsewhere.

We will confirm it with the color map by district, that offers us a clearer vision, since we can appreciate the uniformity that exists in the distribution of housing prices in each one of them. The darker colors represent a higher average price, while the lighter colors indicate the opposite. In this way, the city center and its surroundings have a higher price than the rest of the districts:

Log(Price) by District

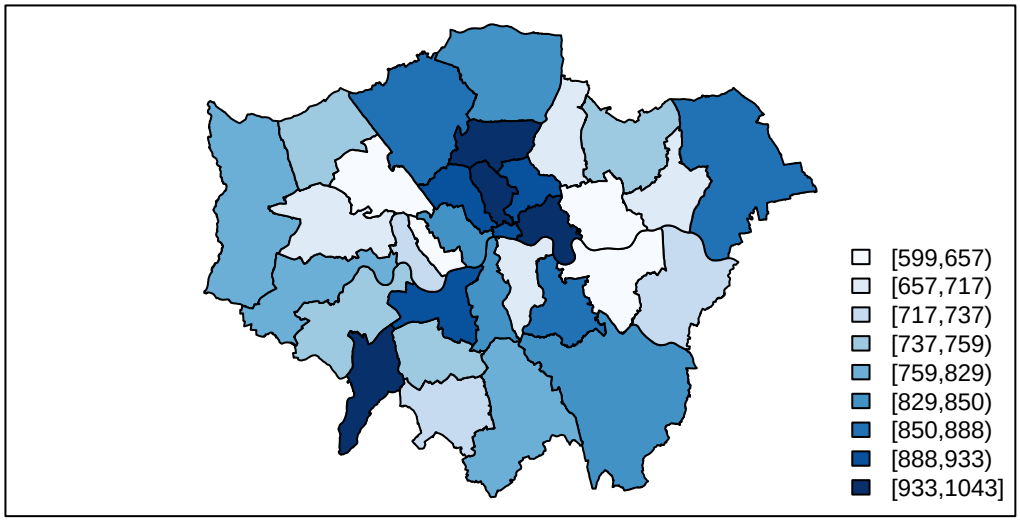




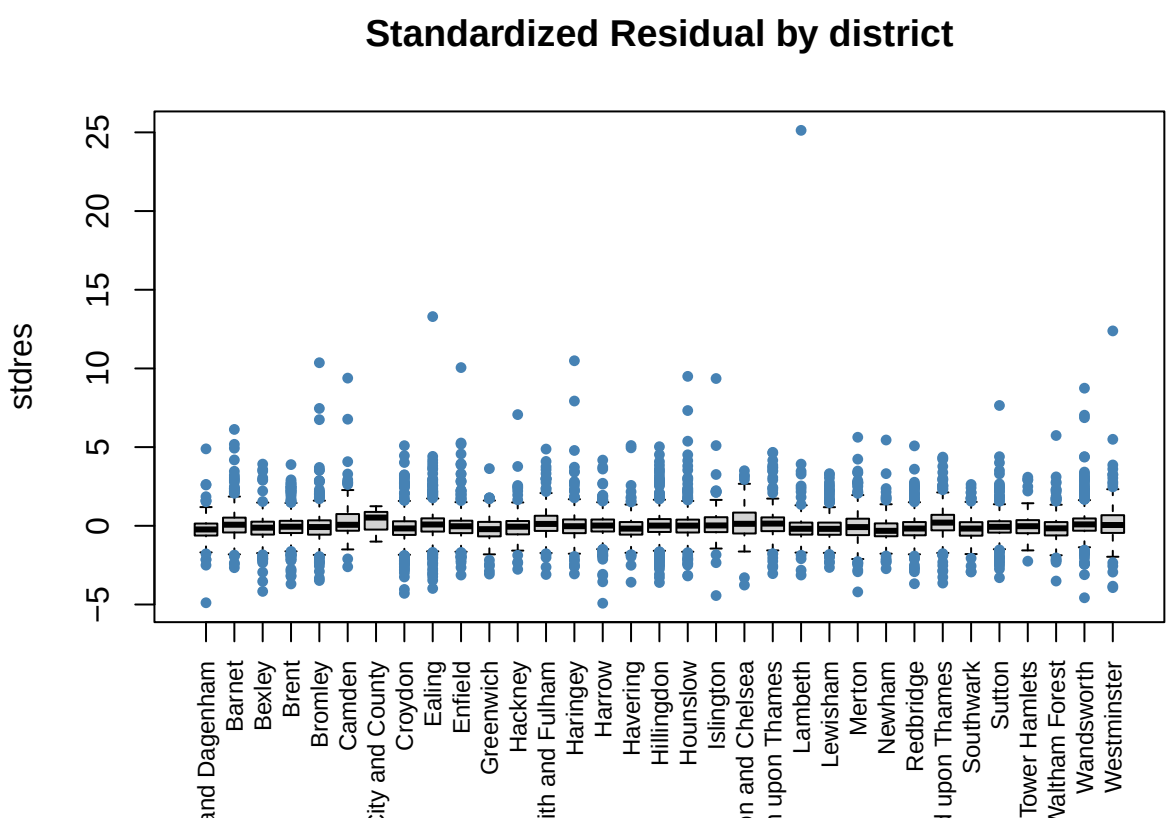
In the case of the median variables in each of the districts, we can see that the median surface area of properties undergoes slight changes from district to district.

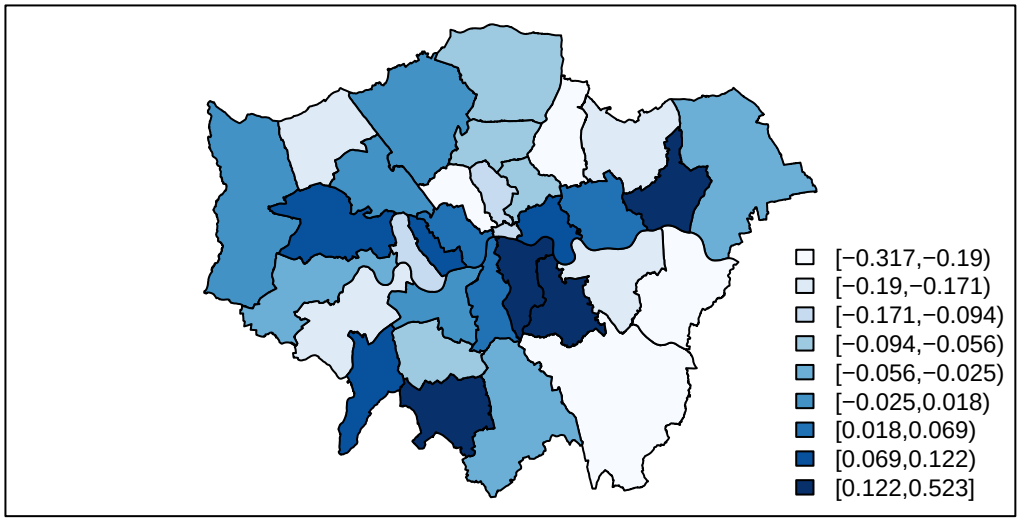
Although the box plot tells us that, except for London City, the median values are not very different, the district map does show the difference: as we get closer to the city center, the median area decreases , with which we can think that the local model fits better:

Floor Area by District



If we look at the following plot, the median residuals by district indicate that there are no relevant changes in these values. On the map, the residual median is slightly more significant in the northwest area of the map compared to the rest of the areas.


Standardized Residual by district
stdres
−5
0
5
10
15
20
25
and Dagenham
Barnet
Bexley
Brent
Bromley
Camden
City and County
Croydon
Ealing
Enfield
Greenwich
Hackney
ith and Fulham
Haringey
Harrow
Havering
Hillingdon
Hounslow
Islington
n and Chelsea
upon Thames
Lambeth
Lewisham
Merton
Newham
Redbridge
upon Thames
Southwark
Sutton
Tower Hamlets
Waltham Forest
Wandsworth
Westminster



7. Final conclusions

We have developed, at the beginning of this project, the aspects that are considered decisive as drivers of the increase in the price of housing in the city of London. For our part, we have tried to delve even deeper by analyzing the tangible factors that determine the price of housing, in order to be able to consider, from an objective point of view, all the variables that influence the price of housing. And of course, we believe we have answered the questions we asked ourselves at the beginning.

We believe, after finishing this project, that there are measures and tools to be able to affirm what the market trend will be according to a series of variables and conditions, as we have been able to appreciate. And we can ensure that the data is real (see Bibliography) and is available to anyone who wishes to dedicate time and effort to it.

8. Bibliography

The *economic information* mentioned at the beginning of this project has been obtained from these sources:

https://www.economicshelp.org
https://www.london.gov.uk/sites/default/files/house-prices-in-london.pdf
https://www.progressiveproperty.co.uk
https://www.reuters.com
https://www.trustforlondon.org.uk

The dataset is a subset of mortgage records and other variables for the area known as Greater London and has been extracted from London Datastore. It is released under UK Open Government License. All the information from this

project, can be found here:

London Boroughs (Districts) Map
https://data.london.gov.uk/download/london_boroughs/9502cdec-5df0-46e3-8aa1-2b5c5233a31f/London_Boroughs.gpkg

Dataset
https://data.london.gov.uk/dataset/uk-house-price-index
https://data.london.gov.uk/dataset/number-and-density-of-dwellings-by-borough
https://data.london.gov.uk/dataset/jobs-and-job-density-borough
https://data.london.gov.uk/dataset/local-authority-average-rents
https://data.london.gov.uk/dataset/ratio-house-prices-earnings-borough
https://data.london.gov.uk/dataset/average-house-prices
https://www.gov.uk/search-property-information-land-registry