**Property Price Estimation in greater london**

**Author:** Pratik Budhdeo

Table of Contents

[1. Introduction: 3](#_Toc8942936)

[2. Understanding the Data Set: 4](#_Toc8942937)

[3. Correlation Analysis: 5](#_Toc8942938)

[4. Outlier Analysis: 7](#_Toc8942939)

[A). Continuous Predictors 7](#_Toc8942940)

[B). Categorical Predictors 8](#_Toc8942941)

[5. Predictor Significance: 11](#_Toc8942942)

[6. Linear Regression: 12](#_Toc8942943)

[7. Random Forest: 14](#_Toc8942944)

[8. Spatial Analysis: 16](#_Toc8942945)

[9. Spatial Visualization: 21](#_Toc8942946)

[10. Summary: 23](#_Toc8942947)

# 1. Introduction:

The purpose of the project is to analyse and find the most decisive ***determinants of housing prices in Greater London area***. Statistical analysis plays a vital role in modelling the price of a property as a function of its attributes, and some of the characteristics of neighbourhood.

Using exploratory analyses and different Machine Learning techniques in R, we need to establish the relationships between several predictors variables and a response variable (Price) in this project. The results would help to better understand the estimates, spatial variation and predict outliers. The data provided is over 25 years old, and the sales came at the end of a decade when property prices had started to rise quickly following the deregulation of financial sector under Margaret Thatcher’s government and the passing of legislation allowing council tenants to buy their own council houses.

Some of the most commonly used regressions techniques are Linear, Logistic, Ridge, Lasso, and Random Forest. After evaluating multiple models and comparing them, best model is evaluated graphically to visualize significant predictors and its impact on housing price estimation.

# 2. Understanding the Data Set:

The dataset provided for the analysis are anonymized, coded to unit postcode level in UK for the houses built between in the *1918 till 1989*. There are total **12536 rows**, each representing one property having **31 attributes**. The response variable we want to predict is “***Purprice***”. The exploratory variables include three categories: -

1. **Spatial** – *like* Easting, Northing
2. **House size and facilities** – *like* Bld60s, TypDetch
3. **Neighbourhood characteristics** – *like* CarspP, RetiPct



***Figure 1: Variable in Dataset***

The above figure 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.

***Dates:*** Represent the time period in which the property was constructed. The omitted category is built pre-1914.

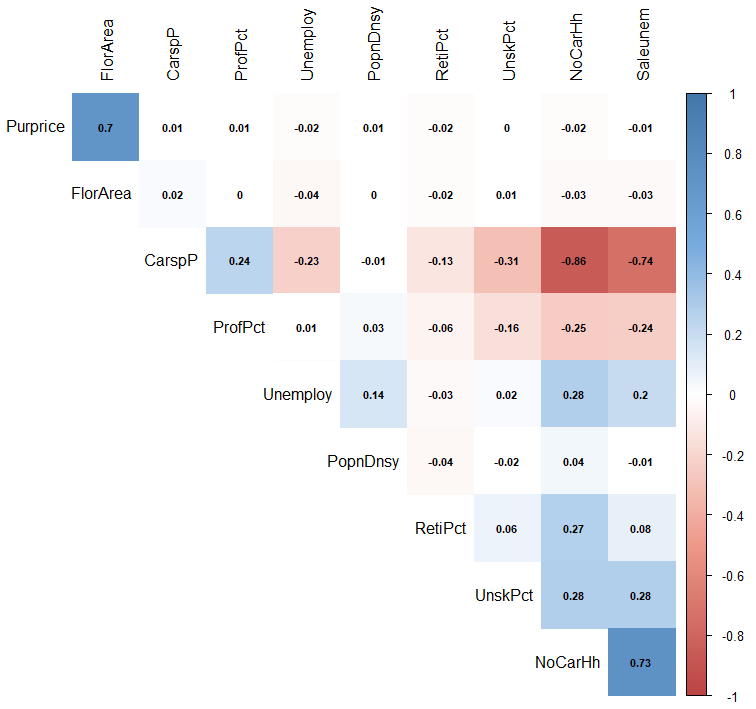
***Tenure****:* Represents the type of building. The omitted category is bungalow.

***Garage***: Omitted category is NO Garage.

***Bedrooms****:* Omitted category is One Bedroom.

# 3. Correlation Analysis:

Before analyse model, it is important to identify the relationship between the predictors so that we can prevent any high influence predictor to deviate the results if it is not significant. Best way is to **minimize the collinearity** between the variables to reduce the influence. In below graph, darker shades of blue and red conveys the high positive and negative association.

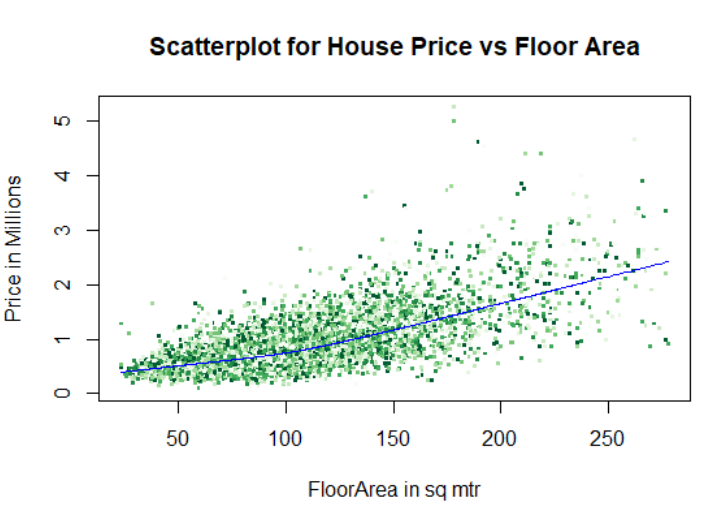


High correlation pairs (above 0.7) are as follows: -

* Response Purprice with FlorArea (+0.7)
* Saleunem with NoCarHh (+0.73) and CarspP (-0.74)
* NoCarHh with CarspP (-0.86)

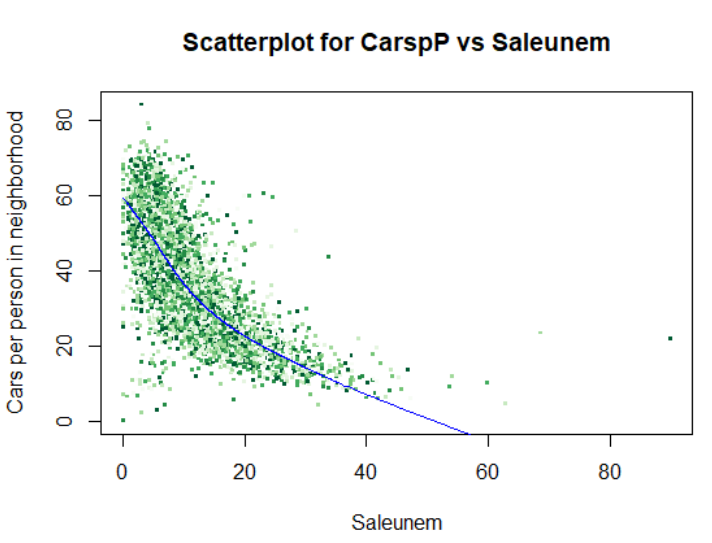
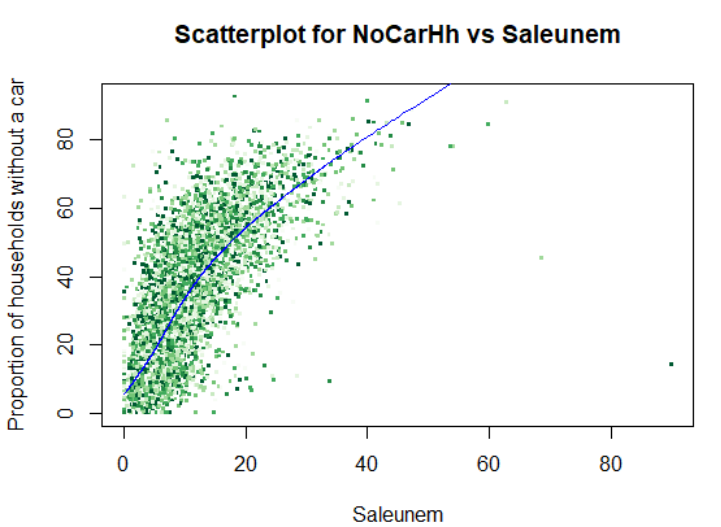
***Fig 2: Correlation plot for numerical variables***

To further understand the relationship between property price and floor area, we create a **scatterplot** **with lowess line which allows nonlinear boundary** that helps to identify exactly how the price is increasing with increase in Floor Area.



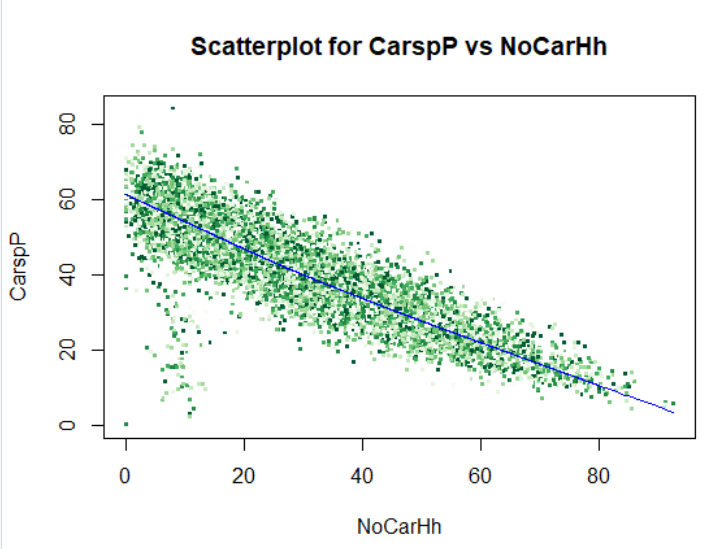
The rate of increase in price is higher for apartments with floor area greater than 150 sq. mt., compared to a slow rise for house below this range.

***Figure 3.1: House Price vs Floor Area***



***Figure 3.2: NoCarHh vs Saleunem and CarspP vs Saleunem***

Saleunem is increasing with increase in NoCarHh and decreasing with increase in CarspP.



***Figure 3.3: CarspP vs NoCarHh***

CarspP Cars per person in neighborhood decrease if Proportion of households without a car increase. These two variables are actually measuring the same thing. One of the variables can be removed from the analysis.

# 4. Outlier Analysis:

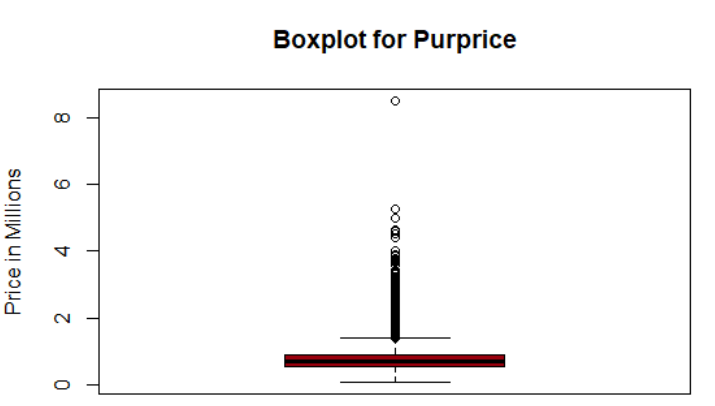
**Boxplot** is a great tool to detect outliers and inspect them visually. It is a standard method of demonstrating the distribution of data based on a five-number summary (“minimum”, Q1- first quartile, median, Q3 - third quartile, and “maximum”).

These plots help in understanding if the data has uniformity. If it is non-uniform, we can apply transformation techniques to scale them so equal weight is emphasized on variables that are otherwise of smaller range. This makes models more interpretable.

## A). Continuous Predictors

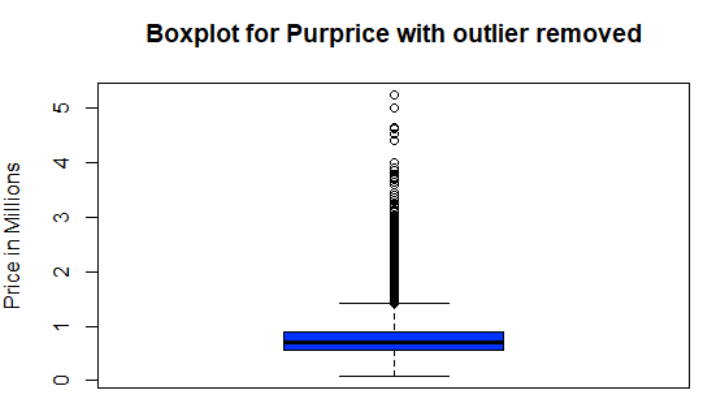
**Using linear regression, we found that UnskPct, RetiPct, Saleunem, Unemploy, PopnDnsy are not important (P-value > 0.05), hence have been removed from analysis and a new dataset is created without these predictors.**

Boxplot for property price.



All the properties are below 6 Million GBP except one, that is above 8 Million displaying. We can remove this outlier and re-visualize the graph.

***Figure 4.1: Boxplot with outlier***



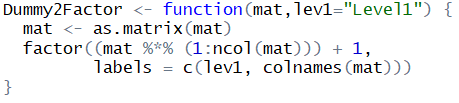
In the wake of dispensing with the outlier, we see that the value of remaining houses is in range 0-6 million. So now on the off chance that we fit the information, the outcome will be precise.

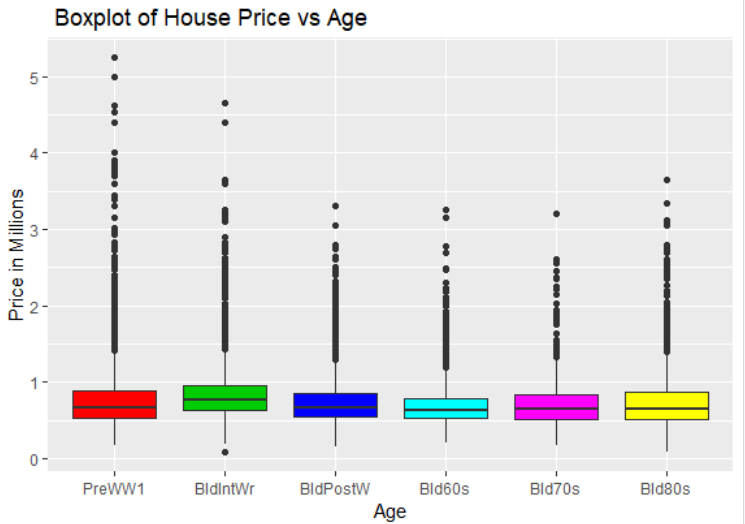
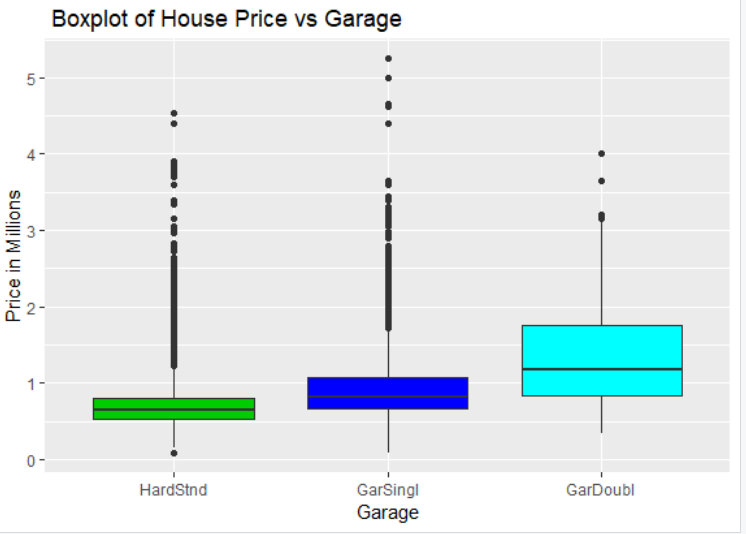
***Figure 4.2: Boxplot without outlier***

## B). Categorical Predictors

In this dataset, some of the Categorical predictors are having just one column like CenHeat and NewPropD. While some predictors such as such as age, type, number of garages, bedrooms are represented as dummy variables which belong to same category, we can merge them treat as a single categorical variable.

This can be done using a simple function “***Dummy2Factor***” as shown below. This takes in the columns as matrix and then converts into a single column.

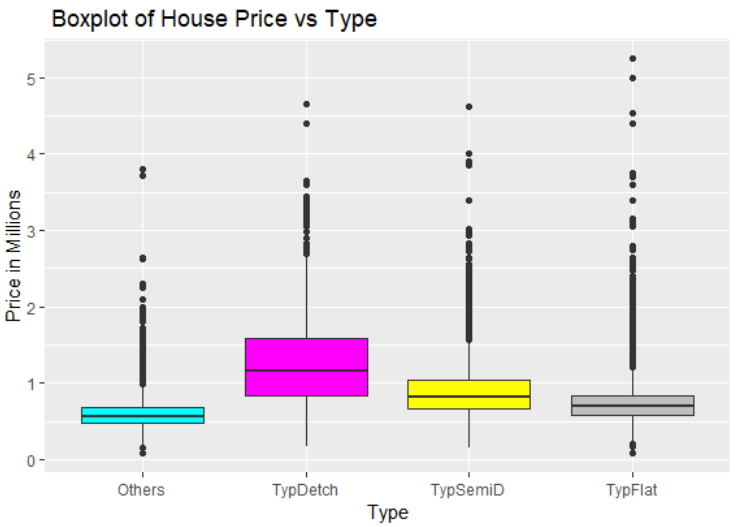
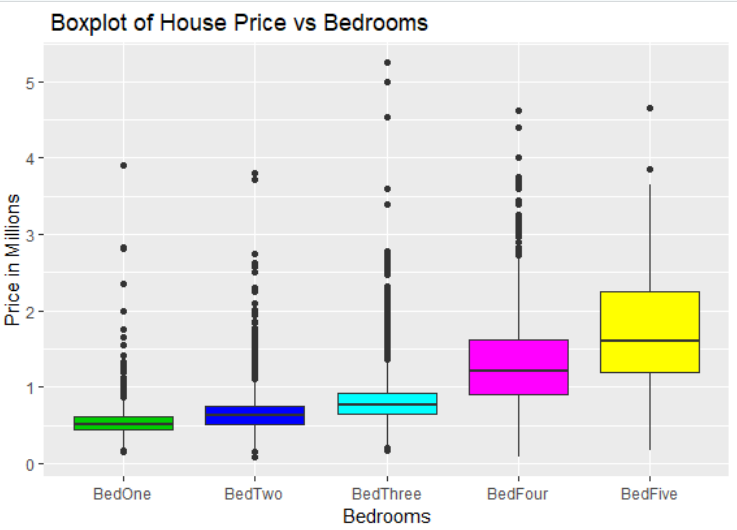


***Figure: 4.3 Boxplots for Age and Garage***

***In Figure 4.3***, we observe below findings: -

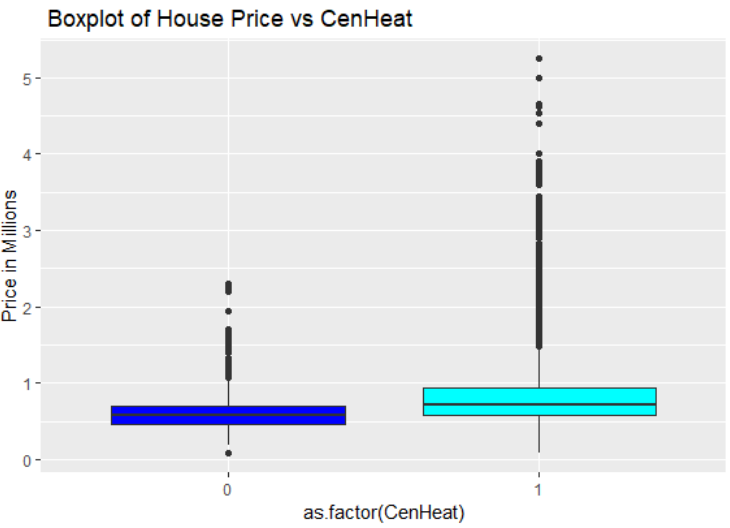
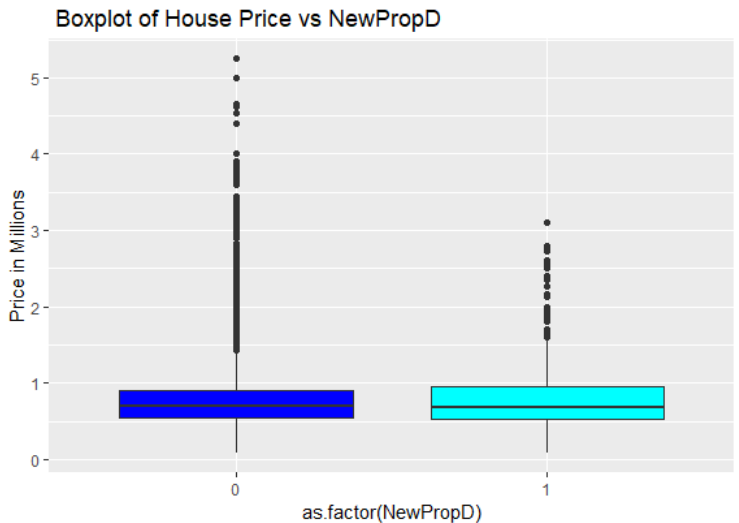
* **Purprice vs Age** - BldIntWr has highest median price amongst all categories, but PreWW1 has highest number of houses in price range above 3 Million GBP. The overall distribution is optimized with some outliers in all categories. All the remaining aged buildings have approximately the same median values.
* **Purprice vs Garage** - We observe that GarDoubl has maximum spread in the inter-quartile range and Hardstnd (No Garage) has the least value i.e. majority houses in the dataset have double garages. Also, GarDoubl has a very high median value, suggesting it is the only category where median housing prices are above 1 Million GBP, while remaining two categories Hardstnd and GarSingl have median prices well below 0.9 Million GBP. There are a few outliners in all the categories which is not unusual, as they might have other better features that are not explained only by this predictor.

***Figure: 4.4 Boxplots for Type and Bedrooms***

***In Figure 4.4***, we observe below findings: -

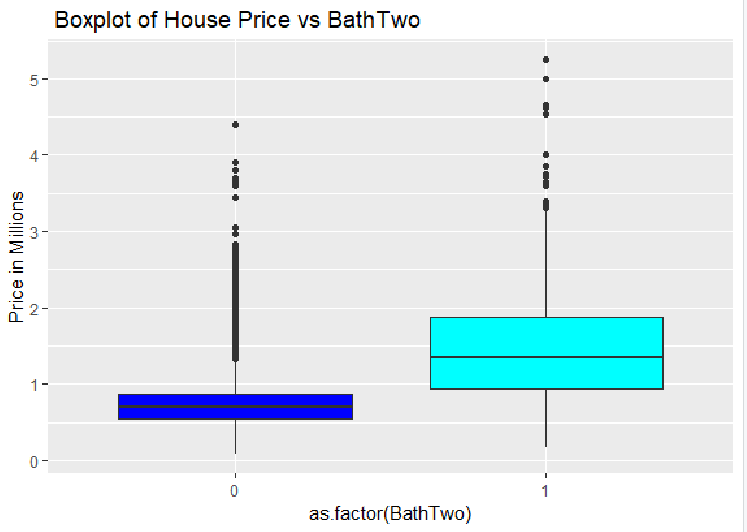
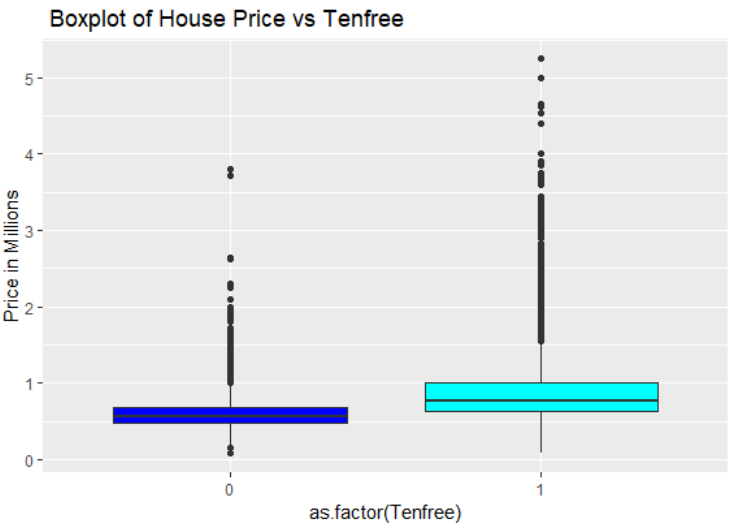
* **Purprice vs Type** – This plot suggest that people are willing to pay higher price for TypDetch (Fully Detached) houses as compared to Semi-Detached or Flats. Interesting thing to note here is that some of the Flats have price values in the range of over 3.5 Million GBP shown as outliers in the 4th column depicted by grey colour, and have the highest overall spread in terms of Price. This means it would be difficult to estimate price of flats if other predictors are not explaining this variability in the housing price. All categories except for TypDetch have median price less than 1 Million GBP.
* **Purprice vs Bedrooms** – As Expected, there is increase in median price with increase in number of bedrooms, five bedrooms houses have highest price range with median above 1.5 Million GBP. Some outliers for three-bedroom houses have value greater than five-bedroom houses above 3.5 Million GBP. It might be because of their locations or the other luxury like central heating, garages, parking lots, two bathrooms, etc.

***Figure: 4.5 Boxplots for CenHeat and NewPropD***

***In Figure 4.5***, we observe below findings: -

* **Purprice vs CenHeat** - CenHeat being a categorical variable with values 0 and 1, it is expected that the house with the central heating center are highly priced than the ones without it. Also keeping the location and weather condition of UK in mind, people would want to have a housed equipped with central heating due to extreme cold weathers during winter. All the houses without CenHeat have price below 2.5 Million GBP.
* **Purprice vs NewPropD** – It is unexpected to see that old houses have higher median compared to the newly made houses. There are 12079 old properties and only 456 new ones, hence the data is unbalanced and it’s not fair to compare these categories without scaling. This tells us that during the time period when the data was collected, there were very few new houses in UK and the picture might have changed significantly now.

***Figure: 4.6 Boxplots for and BathTwo and Tenfree***

***In Figure 4.6***, we observe below findings: -

* **Purprice vs BathTwo** - Houses with two (or more) bathrooms have median of 1.4 Million GBP. 11860 houses have only 1 Bathroom and 675 have two or more, again the data is unbalanced and it’s not fair to compare these categories without scaling. Houses with 1 bathroom are priced lower with median price around 0.8 Million GBP.
* **Purprice vs TenFree** - Freehold indicator in a property means that you own the land and buildings (if any) outright. So, considering this we see that as expected; properties with indicator as 1 have a higher median price than with indicator 0.

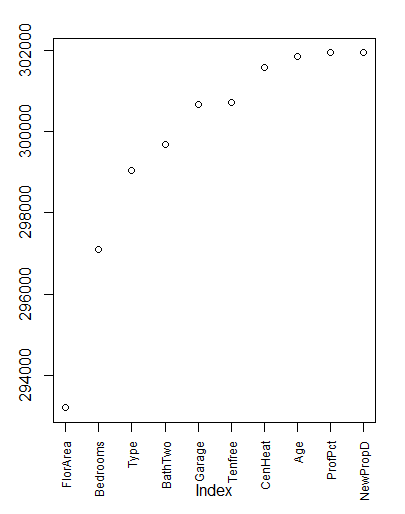
# 5. Predictor Significance:

Identification of true relation between predictors and response requires a model estimation technique. In this dataset, the response variable is continuous so regression setting is favorable for modelling. We can perform regression of Purpricewith each predictor and select appropriate model lowest AIC value, which will give us the closest model to the unknown actual model.

On analyzing the AICs from all the models, the best model with combination of predictors having least AIC is selected. The AICs of the single predictor model are contrasted with Floor Area. In this way, this is the nearest model to the unknown true model.

|  |  |
| --- | --- |
| **Predictors** | **AIC** |
| NewPropD | 301930 |
| ProfPct | 301929 |
| Age | 301833 |
| CenHeat | 301562 |
| Tenfree | 300712 |
| Garage | 300656 |
| BathTwo | 299668 |
| Type | 299032 |
| Bedrooms | 297087 |
| FlorArea | 293199 |

Above table can be visualized as a graph and influential variables can be derived. Type and Bedrooms are most important predictors after Floor Area and NewPropD, ProfPct and Age are the lease significant variables.

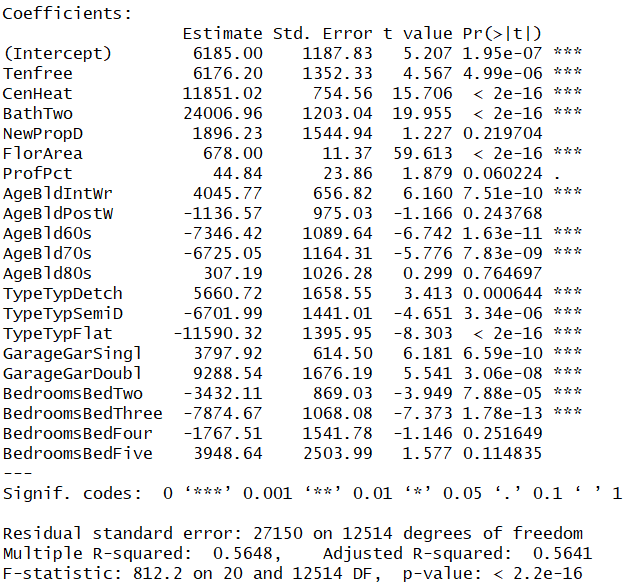


***Figure 5: Index vs AIC plot***

# 6. Linear Regression:

Based on the prior analysis using boxplot we are confident that a relationship exists between the predictors and the response variable “Purprice”, but not sure that if it is linear or non-linear. Initially we will attempt to fit the simplest model with all the predictors and analyse the P-values for significance.

Linear regression helps to derive the relationship between a response and predictor variables (explanatory/independent variables). Easting and Northing are coordinates, hence not considered under this analysis and would be dealt separately under spatial analysis.



***Figure 6.1: Coefficients from Linear model***

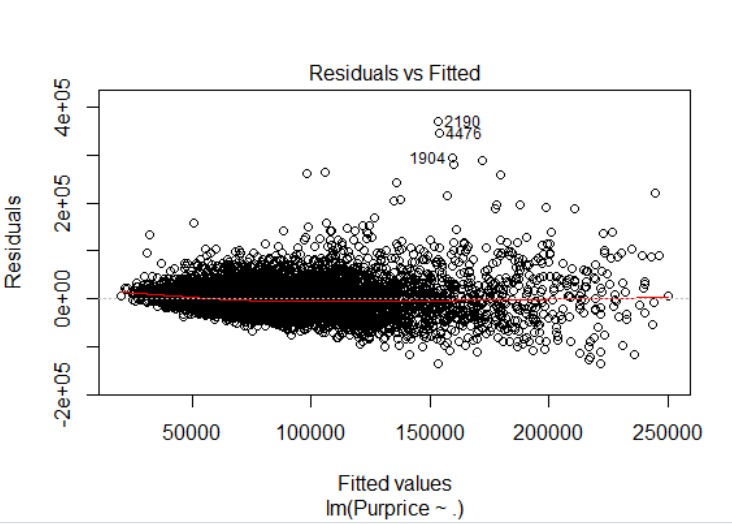
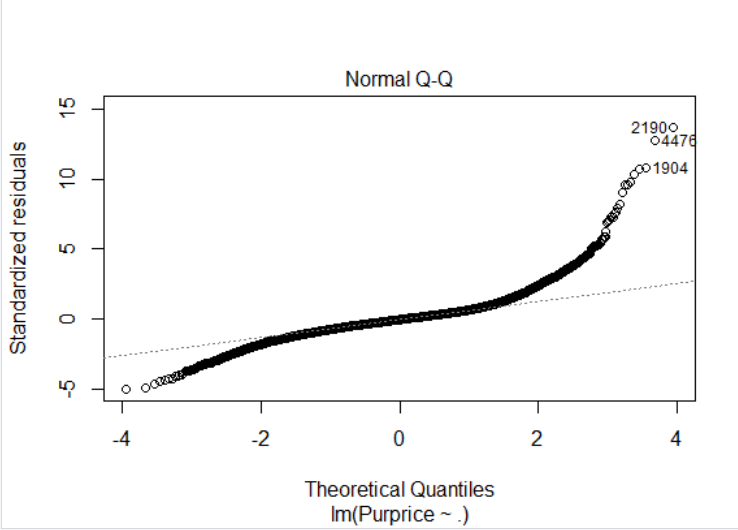
**Significant** predictors with (P < 0.05) having **positive coefficients** are Tenfree**,** CenHeat, BathTwo, FlorArea, BldIntWr, TypDetch, GarSingl and GarDoubl; i.e. Purprice increases when these predictors increase or are present.

**Significant** predictors with (P < 0.05) having **negative coefficients** are Bld60s, Bld70s, TypSemiD, TypFlat, BedTwo and BedThree; i.e. Purprice is negatively associated to these predictors.

Analysing the results based on P-Value (P > 0.05), **insignificant predictors** are NewPropD, ProfPct, BldPostW, Bld80s, BedFour, BedFive.

Important ***conclusions*** from the coefficients: -

* Houses that are detached with Tenfree (=1), centralized heating, two or more bathrooms, garage (1 or 2) will fetch the highest price.
* Having more bedrooms in a property is favorable, but not beyond three.
* Building older than pre-war era fetch more price, this may be due to historic significance.
* House built during 1960 – 1980 have negative effect on the price of the property.
* New property (NewPropD) and higher Proportion of Households with Professional Head (ProfPct) do not play an important role for estimating the property price.

***Figure 6.2: Residuals vs Fitted Figure 6.3: QQ Plot***

***Adjusted R-squared*** is the proportion of variance explained by the linear model, in this case it is **only** **56.48**.

From residuals vs Fitted plot, we see that there is a strong association between the predictors and response, but there is non-constant variance having curved nature. QQ Plot suggest that the residuals are not normally distributed.

Hence the assumptions of a linear model are violated and might not be the best fit. This could be because of the spatial variation of the property price or that the data requires a model with non-linear boundary to distinguish the predictors.

The plot also indicates some probable outlier with obs. no. 1904, 4476 and 2190 which could be same as observed in boxplots earlier.

In our case, supervised learning algorithms are favorable. Some of the techniques are: -

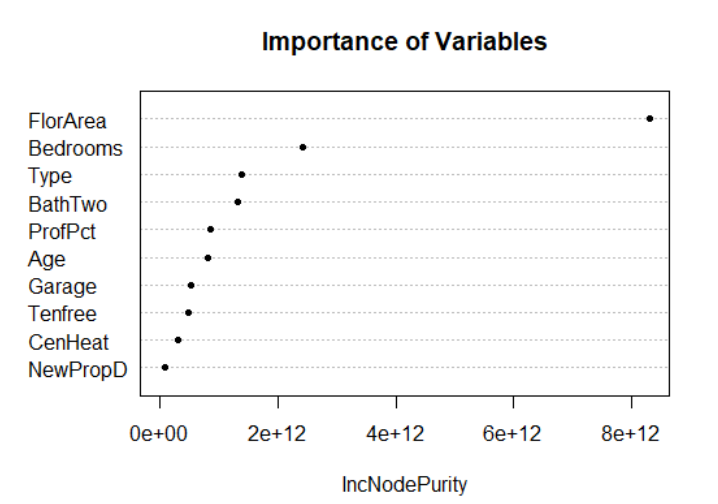
* LDA – Response must be categorical
* K-nearest neighbor – Response must be categorical
* Random Forest – Works with both categorical or continuous predictors.

# 7. Random Forest:

Random forest works by developing a multitude of decision trees at training time and outputting the class or mean prediction (regression) of the individual trees to get a precise and stable prediction.

While training, each tree in a random forest learns from a random sample of the data points. The samples are drawn with substitution, known as bootstrapping. The idea is that by training each tree on different samples, overall, the entire forest will have lower variance but not at the expense of increasing the bias.

With test data, forecasts are made calculating mean predictions of each decision tree.



***Figure 7.1: Importance of Variables plot using Random Forest***

The mean decrease in Gini coefficient is a way to describe how each variable contributes to the homogeneity of the nodes and leaves in the results from Random Forest. Predictors that result in nodes with higher purity have a higher decrease in Gini coefficient.”

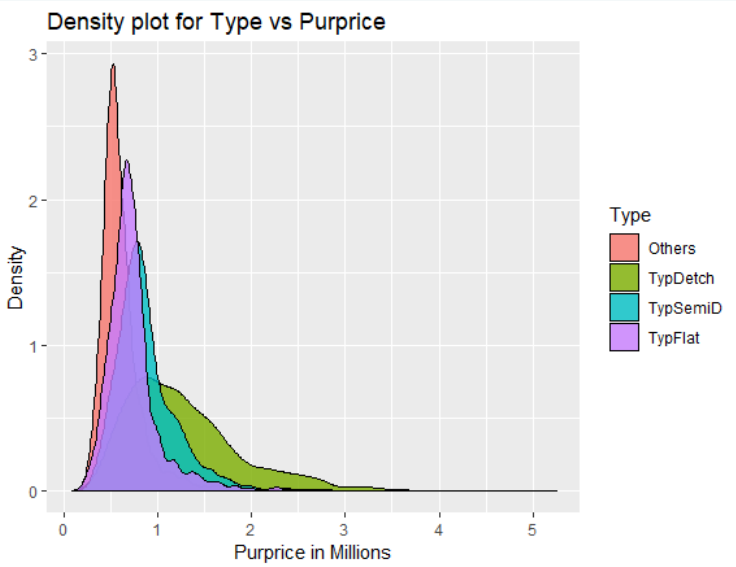
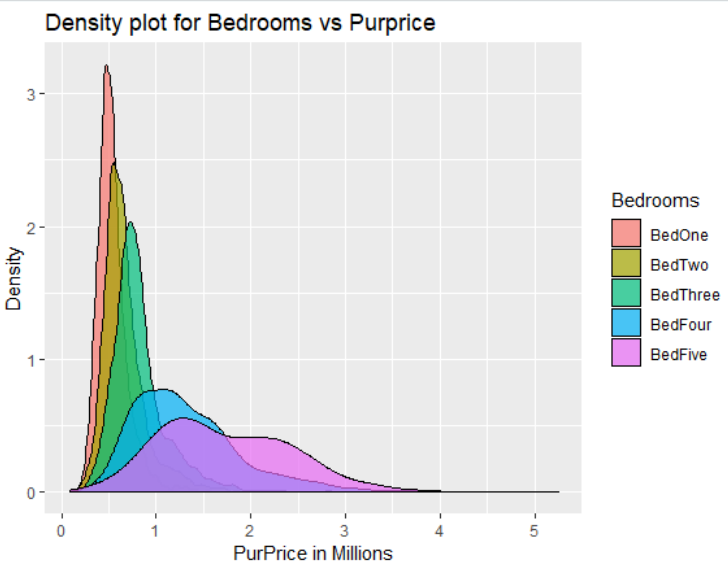
The *importance chart* shows the variables arranged from greatest impact to least impact, from top to bottom. We can utilize these to find out most important predictor variables.

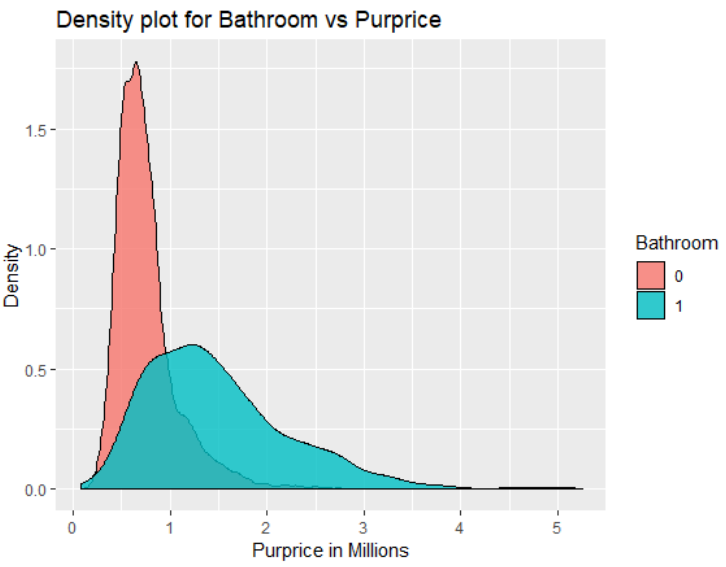
Top 4 important predictors are: -

1. FlorArea
2. Bedrooms
3. Type
4. BathTwo

***Adjusted R-squared*** by this model is **increased to 58.43**, so it’s a comparatively better model than the linear regression model.

It’s interesting to notice the contrast here with linear model that Tenfree**,** CenHeat and Garage are not predicted to be significant as they explain only little variability in the data compared to other important predictors. This means they are not significant in predicting the “Purprice”.



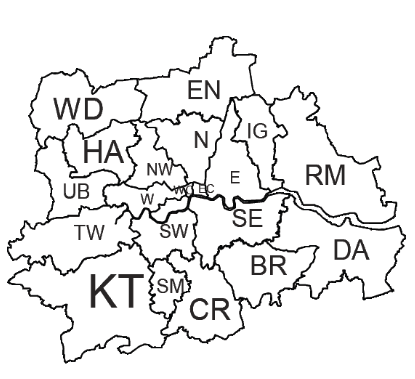
***Figure 7.2: Density plots for significant variables***

# 8. Spatial Analysis:

Spatial analysis helps to examine the influence of locations on the housing price, and interpret the relationships of features in spatial context through overlay.

* **Studying about the Postal code regions in Greater London –**

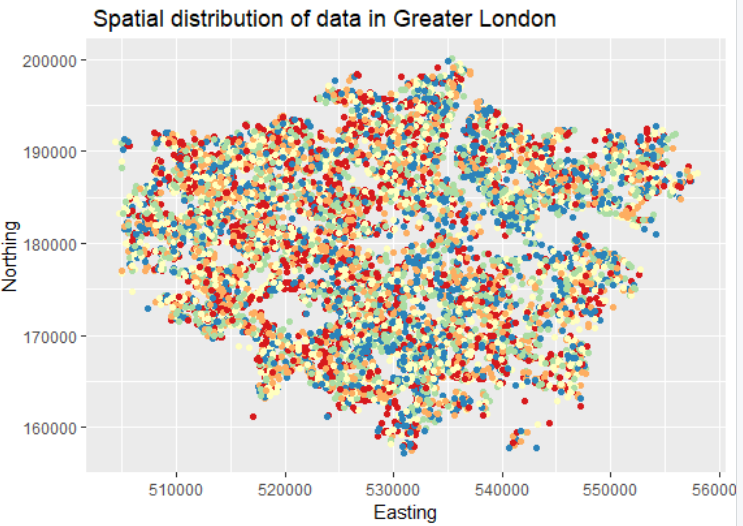
The Greater London area consists **of 21 different postcode** areas. Each area is then divided up into number of smaller districts.



* **Examining the dataset –**

In order for the data to be anonymous in this dataset, first the data were coded to unit postcode level as explained above. Then the grid references were further spatially jittered, keeping the geographical aspect in mind.

* **Visualization raw data –**



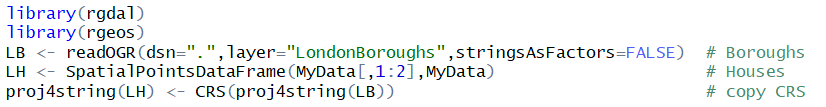
***Figure 8.1: Geographical distribution of houses in Greater London***

From the Graph, we can visualize that the major housing falls in the north ranges between 165000-193000 and east ranges falls in between 510000-550000. Highest density of houses is in the north-west region and least in south-west regions of Greater London.

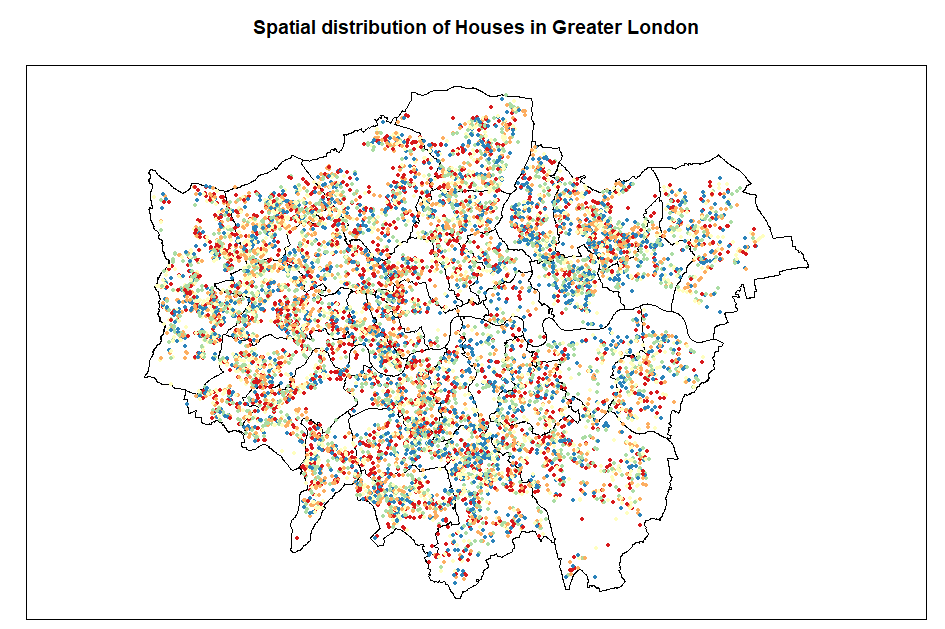
* **Importing Shp files into R and Visualization Data with Map Layout –**

“*LondonBoroughs*” shape file is loaded into R studio for overlay analysis to obtain a demographic knowledge of the model fit. To Load the shapefile, readOGR function is used.

We project the data by converting the easting and northing into same coordinate reference system as that of the shapefile.

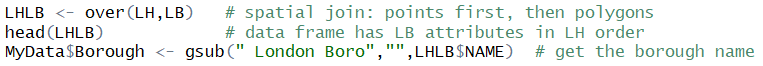


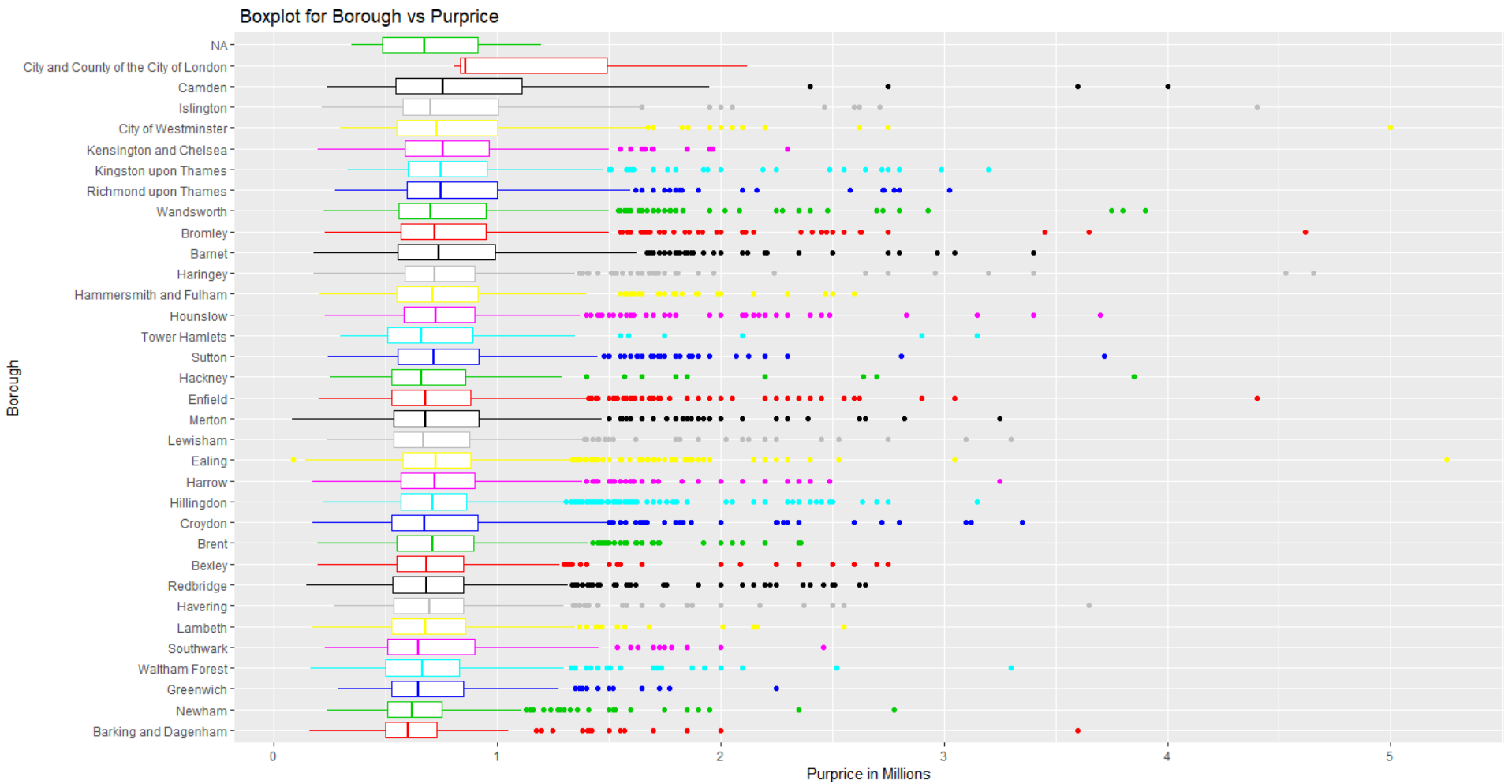
In *Fig 8.2*, we observe that interior areas are densely populated compared to the regions around the edges of the map. This means that there is a **larger population residing near the heart of the city while minimum crowd resides near the exterior boundaries specially in the south-east**.



***Figure 8.2: Houses in Greater London with map layout***

* **Borough Specific Analysis – *“****Over”* function can be used to identify the data points in each borough and then by getting the names of each borough a box plot can be drawn to analyse the distribution of house prices.





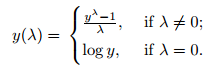
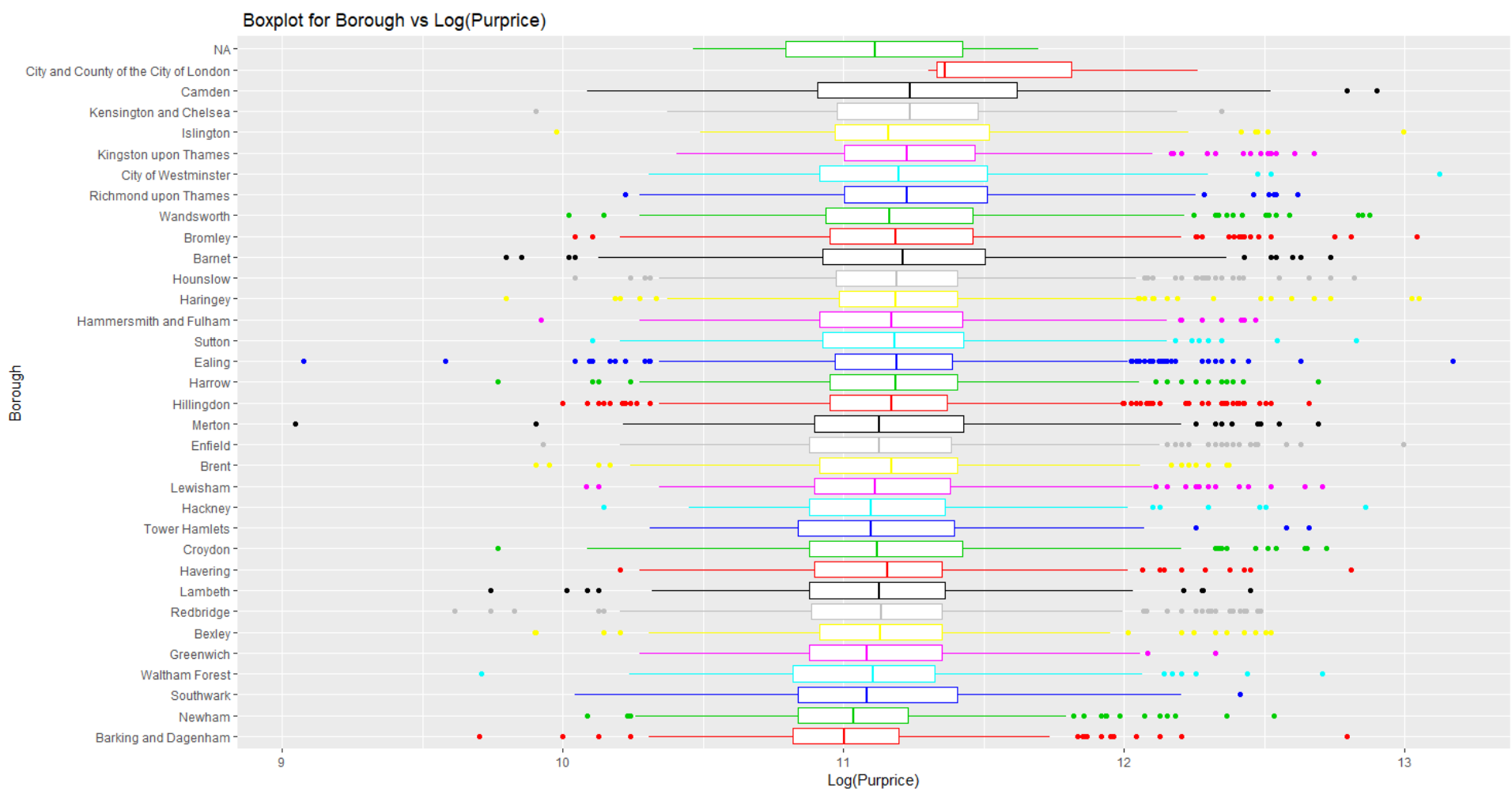
***Figure 8.3: Borough vs Purprice (Sorted Order)***

In *Fig 8.3*, we have represented Purprice of houses in each Borough in a sorted way in terms of the house prices in that region. We derive following conclusions from the graph: -

* City of London is the most expensive area as expected due to its locality with more than 75 % of houses above 1 Million GBP.
* Camden, Islington and Westminster follow next in line after city of London as areas with highest real estate prices. The most expensive house is located in Westminster with a price of 5 Million GBP.
* All the other boroughs have median price in the range of 0.6-0.8 Million GBP, but with high variability/standard deviation with some outliers.
* Barking and Dagenham, Newham, Greenwich and Waltham Forest are the most inexpensive areas in the whole of Greater London with a couple of exceptional properties having price above 2 Million GBP.

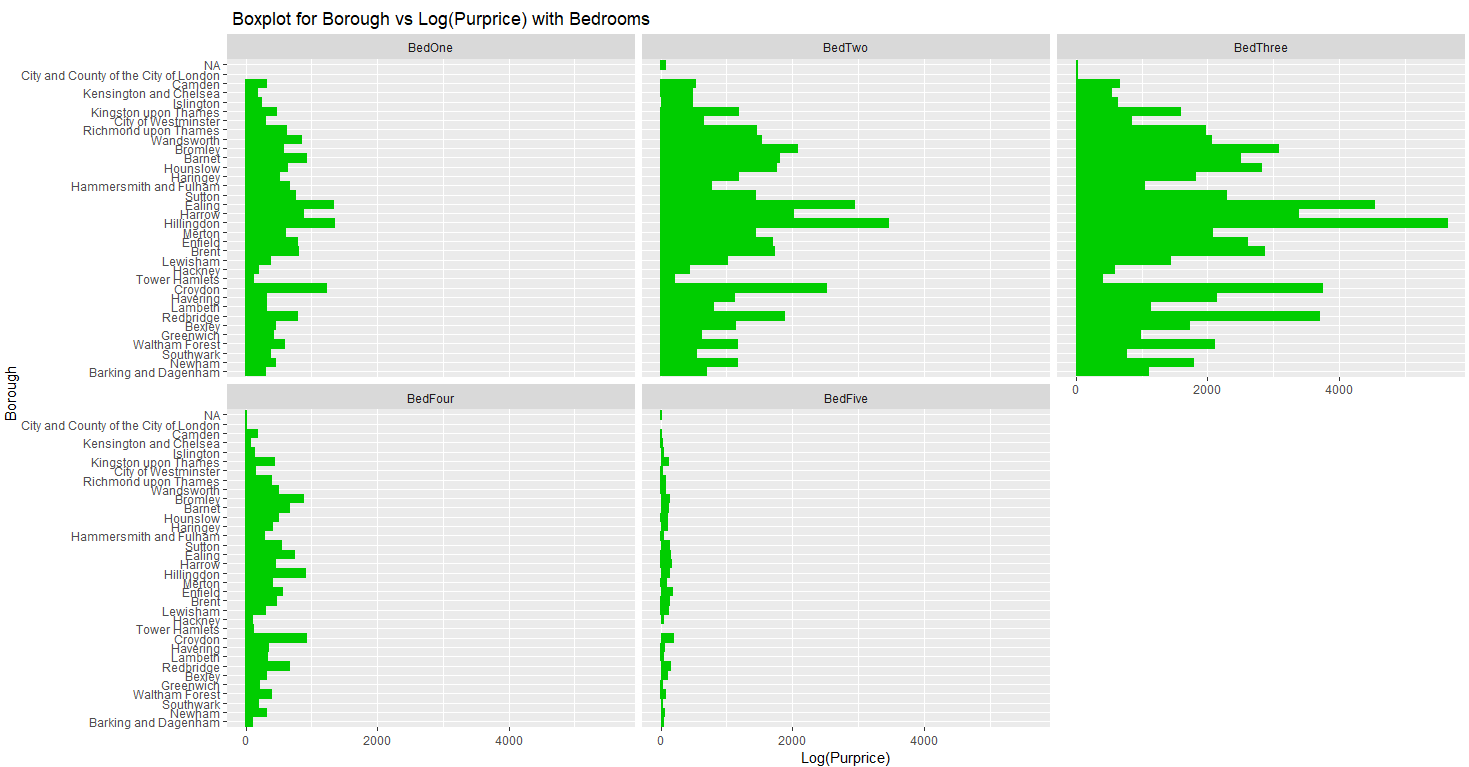
These exceptions in spatial data could very well be explained by other significant non-spatial predictors in the model like FlorArea, Bedrooms, Type, Bathrooms and other amenities.

To reduce the standard deviation in the above plot, log transformation can be applied on the property price. A Box Cox transformation is used to transform non-normal response variables into a normal shape. From the QQ plots it was evident that we need to apply Box-Cox transf. which would help to normalize the residuals.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/07/boxcox-formula-1.png)

***Figure 8.4: Borough vs log (Purprice) (Sorted Order)***

The standard deviation of log transformed purchase price is much less compared to prior plot. Hence, this transformation is suitable for modelling and interpretation.



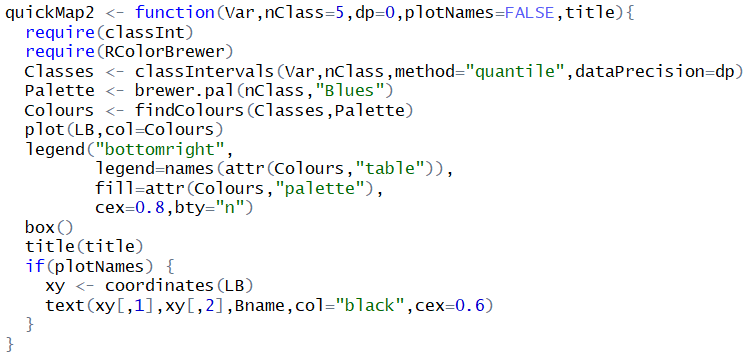
***Figure 8.5: Boxplot for Borough vs log (Purprice) with Bedrooms***

If the price is plotted with respect to each borough based on the number of bedrooms in a property, it is evident that the median house price is different for all the boroughs in each category with high variability. Hence**, borough is significant in estimating the Purprice**.

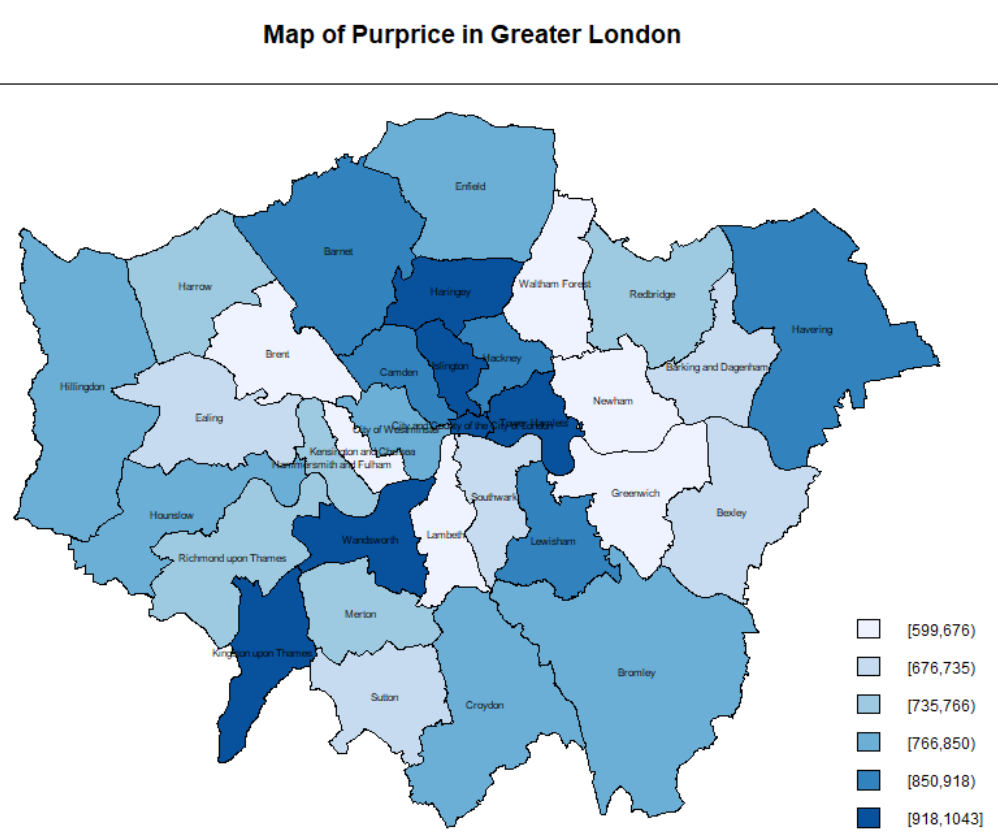
This plot also confirms the results from previous analysis; having more bedrooms increase the price only up to 3 bedrooms, addition of more bedroom beyond 3 do not add significant value to the house.

# 9. Spatial Visualization:

Creating borough specific models for Purprice with FlorArea and storing the coefficients in “results” variable will help us to visualize them on a map. A simple function “quickMap2” to create borough level plots with Legends and title has been written. The class interval selected for this function is 5.



Firstly, we will map the coefficient values for each borough from the model including FlorArea and differentiate it based on color modulation. Boroughs with darker shades depict higher values of coefficients and lighter shades confirm coefficients with lesser value.

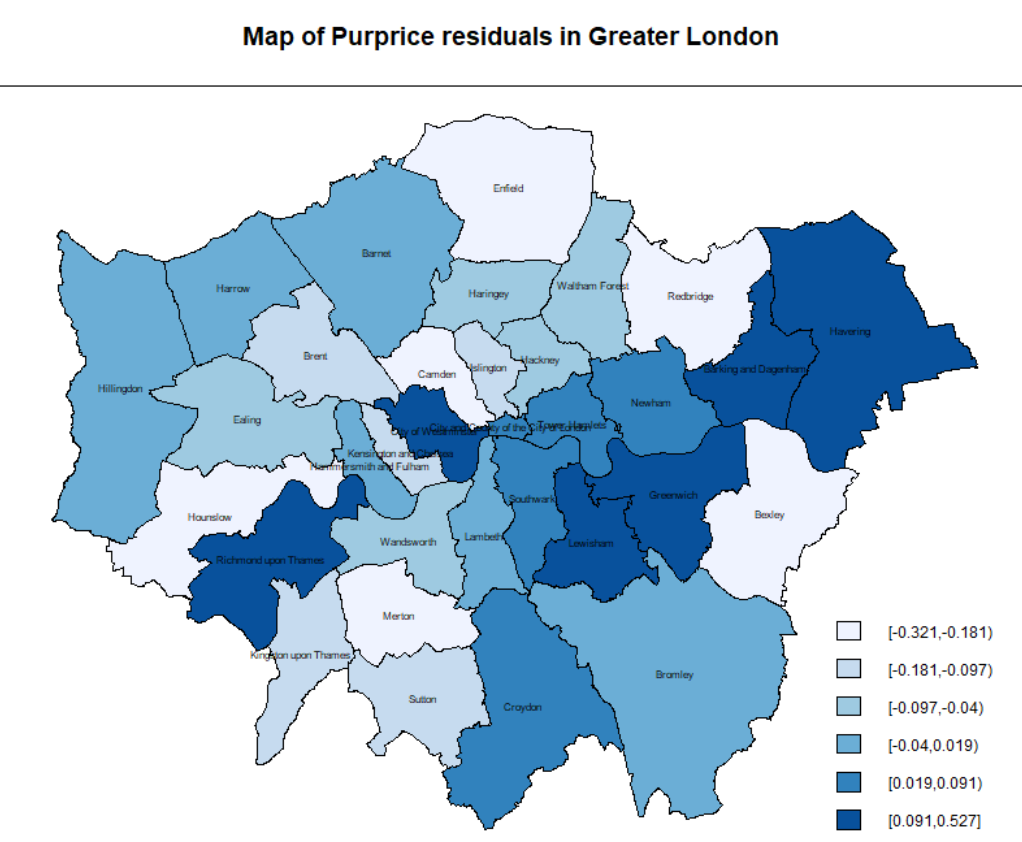


***Figure 9.1: Map of Boroughs using coefficient***

We ***summarize*** the finding below: -

* Kingston upon Thames, Haringey, and Wandsworth, Camden, Islington and Westminster are having maximum coefficient values in the range 903-1043, which means that these places would tend to have a major impact on the housing prices with increase in FlorArea.
* All the remaining boroughs have coefficient in the range of 700-900, they moderately impact the housing prices with FlorArea increase or decrease.
* Brent, Greenwich, Waltham have the least coefficient values below 690 from the model, which means they won’t impact the housing prices as much as the other boroughs with increase in FlorArea.

These **results are similar to the ones obtained using boxplots**.



***Figure 9.2: Map of Boroughs using residuals***

We store the residuals from models for Purprice with FlorArea for each borough using the median values to create the map with residual values. **Darker shaded** regions like Havering and Greenwich have the largest residuals. Bexley, Merton and Hounslow in lighter shades have least residual values. No specific clusters can be formed from the map as the value are changing abruptly with the change in location of boroughs. Overall, the residuals are small and hence the model is a good for prediction for Purprice.

# 10. Summary:

Generally, the **East** End of London is seen as ‘*lower-class’ and ‘poor’*; whilst **West** London is claimed as ‘*high-class’ & ‘quintessentially British’*. Historically, Eastern region has suffered from overpopulation, causing extreme overcrowding, violence and living conditions well below the poverty line. With such a history, it is understandable that the western housing is more expensive that the rest.

The given dataset provided information about the **mortgage records of Greater London area**. Using the given set of predictors, we were successfully able to **predict the housing price variation patterns** and also **understand the significant determinants** of the price fluctuations. Boxplots, Correlation plots, Linear model and Random Forest techniques were used to analyse and build a regression model and visualize the spatial data.

Initially, the variable was analysed separately as continuous and categorical predictors for understanding the collinearity between them. Not all predictors have equal significance to predict Purprice; and this was proved using the results from the models generated that highlighted the significant predictors based on the P-value. The **most important predictor is FlorArea**, and it is directly proportional to the price of a property despite of location. Bigger houses in an unpopular area still costs more that smaller houses in the same area.

A dedicated section is there to analyse the outliers using boxplots and interpret its impact on predictions. Insignificant predictors were identified and removed from the model to reduce the noise as they do not add any value. **Other significant predictors** from the Random Forest model were **Type of house, Bathroom and Bedrooms**. There was a contrasting feature of this model when compared to linear regression that Tenfree**,** CenHeat and Garage were predicted to be insignificant, which were considered significant initially.

There is **significant spatial variation** in the given dataset, i.e. the houses near the centre of London city are priced more compared to ones located near the exterior boundaries of the map. City of London is the most expensive area due its locality with more than 75 % of houses above 1 Million GBP. Camden, Islington and Westminster are also marked as areas with highest real estate prices. All the other boroughs have median price in the range of 0.6-0.8 Million GBP, but with high variability/standard deviation with some outliers. Barking and Dagenham, Newham, Greenwich and Waltham Forest are the most inexpensive areas in the whole of Greater London. With the help regression techniques in R, we were able to analyze and visualize data easily. We succeeded in drawing conclusions about the housing valuation across all boroughs in Greater London.