**PROJECT COVER SHEET**

**Project Title:** “Housing Project”

**Group Members:**

1. **Kumari Sundaram Shailbala – 19250680**
2. **Poojaanjali Vasu – 19251512**
3. **Sangamithra Varadarajan – 19251252**

**Submission Date:** 15th May 2020

**Course Title:** NCG612[A] – Case Studies in Data Science and Analytics

**Lecturer** **:** Project Prof. Charlton, M., Prof. Brunsdon, C

All group members have read and agreed to the final version of all documents.

A close up of a logo

Description automatically generated

**Signatures:**

1. Kumari Sundaram Shailbala
2. Poojaanjali Vasu
3. Sangamithra Varadarajan

**NCG612 - Housing Project**

**Introduction**

Like the saying “Change is the only constant in life”, in the notion of civilization, contemporary living spaces with easy accessibility and smart homes are reviving the basic shelters day by day. Many Land brokering agencies, real estate investors, Buyers are interested to understand the inflation in housing prices depending upon multiple factors. The prediction not only helps in making wise decisions before buying houses, but it also helps in understanding the rise in prices of amenities around the properties. This report seeks to assess the best category of predictor variables used to accurately predict the housing price in London. It also explores a variety of machine learning algorithms to build a suitable model for prediction, using R software. Multiple Regression with OLS statistical methods, Random Forest and GWR(Geographically weighted Regression) were used to find a better fit for the model. This project can be helpful for the people who are trying to buy or sell the property as they will be aware of both current and future scenarios.

**Aim and Objective**

The main aim of the project is to find the best suited model with highest accuracy and also find the predictors highly correlated to each other. The project is also concerned about finding the most important determinants in the price of the property. Another aim for the model is perform a price variation model building.

This report follows a typical data analytics lifecycle. Hyperlinks below can help to navigate to the respective sub-section.

1. [Data Preparation](#_Data_Preparation)
2. [Descriptive Analysis](#_Descriptive_Analysis:)
3. [Predictive Modelling](#_Predictive_Modelling)
4. [Model Validation](#_Model_Validation)
5. [Conclusions](#_Conclusions)

## Data Preparation

The dataset used here is a subdivision of anonymized mortgage records for Greater London and the grid references are spatially jittered. There are 12536 observations and 31 variables of numerical and categorical data. On validating the dataset there are no missing values hence all the observations are considered for analysis. The variables are organised into several groups, and for the most part, represent the levels in a categorical variable expanded into a dummy (0/1) variables. These Dummy variables are factored in order to get a more useful interpretation out of it. So, the predictor variables: BldIntWr, BldPostW, Bld60s, Bld70s, Bld80s represent the period in which the house was built. This is combined as **Age**with Multilevel factor, Similarly, TypDetch, TypSemiD, TypFlat representing the property types namely Detached, semi-Detached, Flat or apartment were combined as **Type**, Garage details GarSingl, GarDoubl are factored as **Garage**, BathTwo, BedTwo, BedThree, BedFour, BedFive as **Bedroom**. Age, Type, Garage, and Bedrooms are derived categorical variables added to the prepared dataset(MyData) replacing the related dummy variables in London Dataset. This modification of variables can be useful for the convenient modelling of data. There are no measures of income in the UK census, so some socio-economic variables are dropped from analysis.

## Descriptive Analysis:

Prior to any analysis, the dataset was checked for multicollinearity to see if any predictors are correlated with each other. From the correlation plot, we could see that the predictors: BathTwo which means two Bathroom type house is highly correlated to FlorArea (Floor Area), it can also be seen that Tenure free which is lease or free hold indicators also correlated to the Floor Area (FlorArea). It can also be inferred from the correlation plot that **Floor Area acts as an important predictor for estimating the Purchase Price for the property**. The same can be confirmed from VIF too as all the predictors have low VIF value. The low Variation Inflation Factor indicates that the predictors are not correlated to each to other.

A close up of a device

Description automatically generated

Figure: Correlation Plot for all the variables

It’s been observed that more than 75% of the houses have higher Purchase price than median price range. The purchase price was also considered for more than 600000 Great Britain Pound and the same trend was found for the property which has higher purchase price than 600000 GBP.

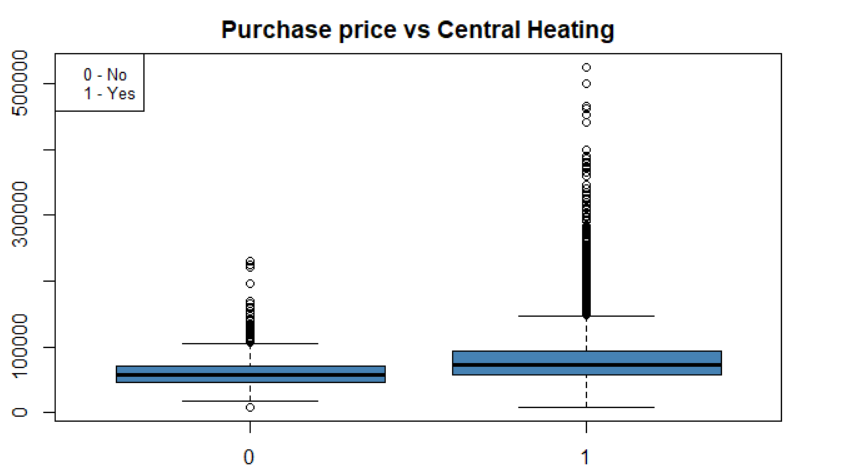
The data was plotted using scatter plot to get more clarity on it. The variables Northing,BldPostW,Bld80s,BedFour,BedFive,NewPropD,NoCarHh,CarspP,ProfPct,UnskPct,RetiPct,Saleunem,Unemploy,PopnDnsy are insignificant and thus removing it.The scatterplot shows that the data had a few suspicious observations for proportion of residents retired (RetiPct) and Unemployed workers (Unemploy). Also, from the scatterplot, it was established that the larger houses which the property with higher Floor Area has higher purchase price so we can say that the variation in purchase price can be considered as a function for variation in the floor size.

## A close up of a map Description automatically generated

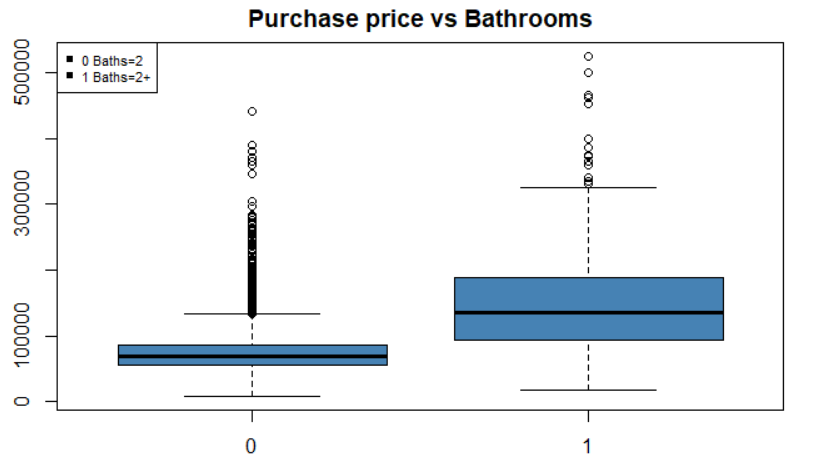
The House Price Prediction csv dataset is loaded in R and we have visualized the house prediction price using box plot and observed that more than 75% of the houses have higher purchase price than the median value as mentioned in data exploration and cleansing. For better analysis, few columns in dataset is converted to factors. The goal of undertaking price variation modelling is model building - which variables represent the tradeoff between adequacy of predictive power versus parsimony. We observe the following while visualizing Property Price with few variables.

The different boxplot was plotted and trends that were observed are as follows:

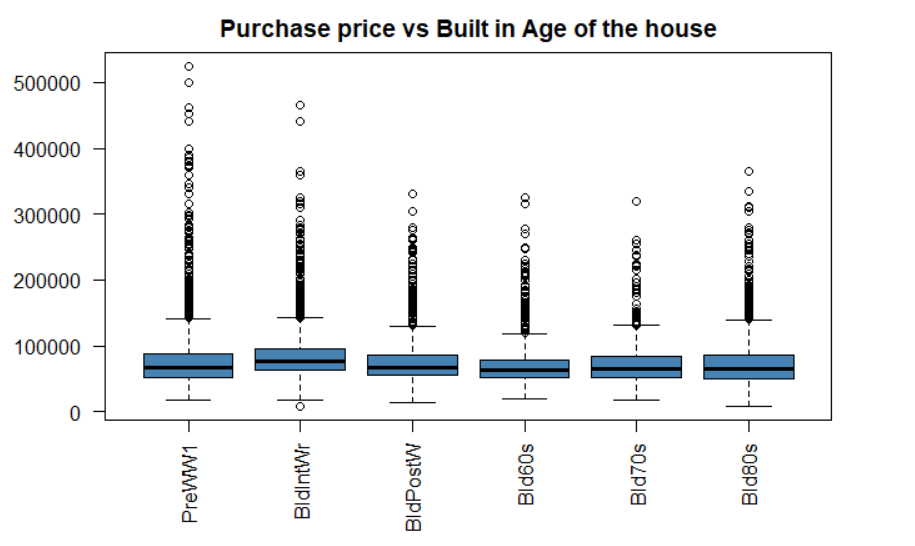
1) For central heating versus Purchase Price plot, if there is no central heating in the house, the price is less and if the house has central heating implanted to it, the price of the property is comparatively more. However, in the both cases, the house price range is above the median value.



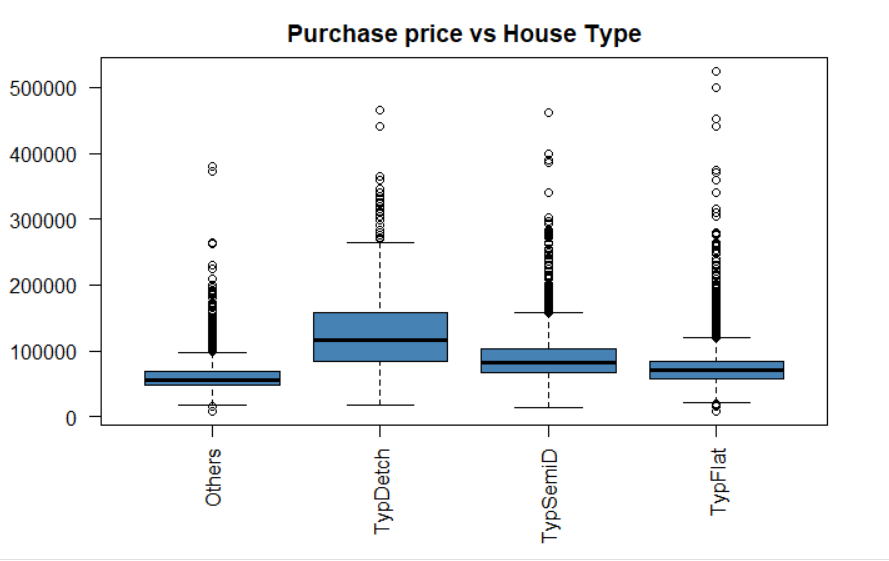
2) Similarly, the buildings with two or more bathrooms have comparatively higher price than the houses with less than two bathrooms. Here also the price range is more than the median value.



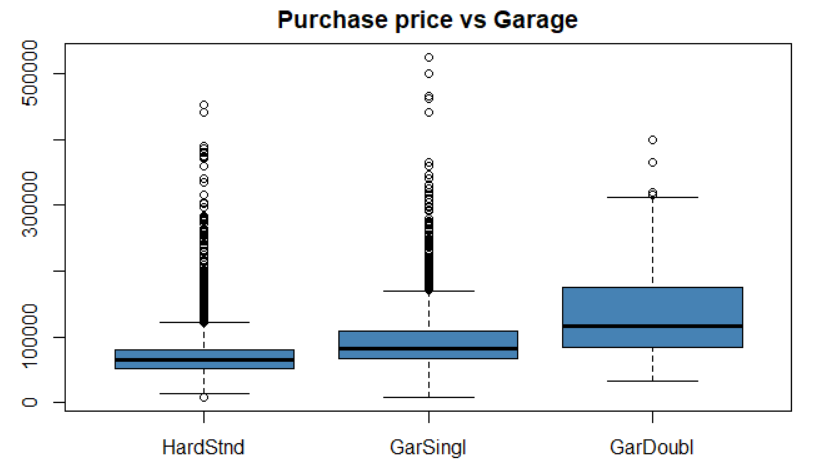
3) The data was collected over 25 years, so when the Age of the building was plotted with the purchase price, there was a pattern before World War and After World War. It has been observed that highest price of houses was in PreWW1 time period which can make sense as there was a housing crisis at that time period, later in 1945- 1979 the price of the houses started to decrease and again there was increase in the prices were observed during 1980-89 time interval. It was observed that the plot of age variable with the purchase price also shows that the price was higher than the median value.



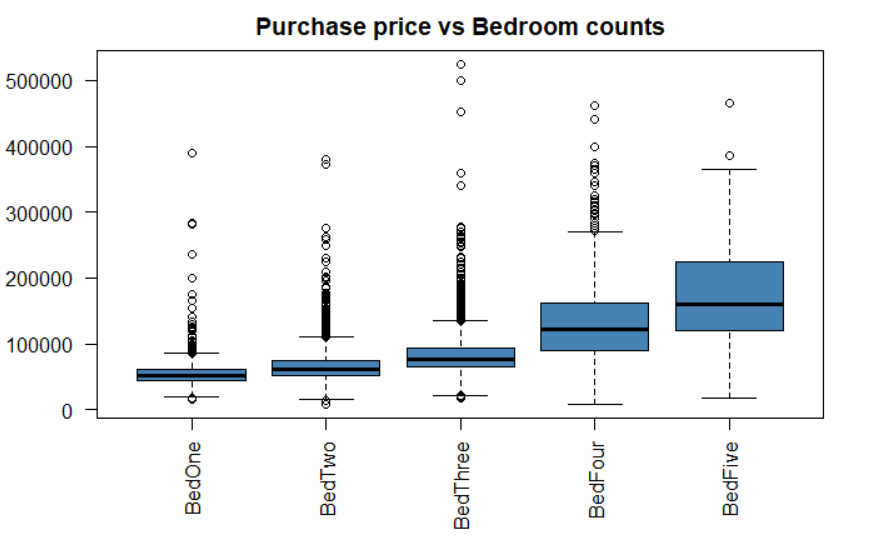
4) When the data was plotted for the price with the Type of the houses, it was observed that the detached property purchase price is comparatively high than the other types of properties. Also, it’s been noted that, the flat type of houses is preferred by the people more than other types, the one among many reasons could be because they are cheaper, and a few people prefer detached houses which might be because they are more expensive than others.



5) When the Purchase price of house was plotted with respect to Garages, it was spotted that the double garaged houses are more expensive than Single Garage and No Garage at all. It was also visible that it was purchased by only few people which can be justified by the fact that these were expensive. Then the single garage houses are comparatively having less price than the double garaged and are frequently purchased than the other types. The houses with no garage are having the least price and not very frequently bought.



6) When the visualization of the Bedrooms with respect to Purchase Price of property was compared, it was clearly inferable that the houses with five bedrooms have the highest purchase price among all the other type of bedrooms which are Single bedroom, double bedrooms, three bedroom and four bedrooms houses. Also, it is deducible that there are a smaller number of buyers in this category which can be justified with the reason that five-bedroom type houses are expensive to afford. The houses with three number of bedrooms have the greater number of buyers as it is easier to afford and cheaper comparatively.

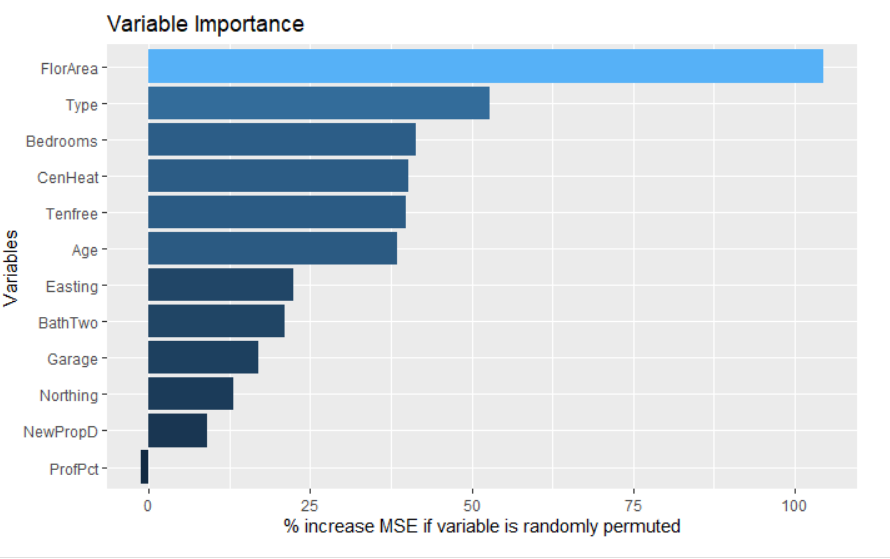


### **Variable Importance**

When Building a Model, it is appropriate to identify the most significant predictors to overcome the problem of multicollinearity, overfitting, underfitting and inaccurate predictions. After preparing the dataset there are 12 predictor variables out of which 2 are spatial co-ordinates Easting and Northing will be used during spatial analysis in model prediction. There are various methods to identify the best predictors for a model. Stepwise Regression and Random forest are used to visualize and determine the significant predictors in this report.

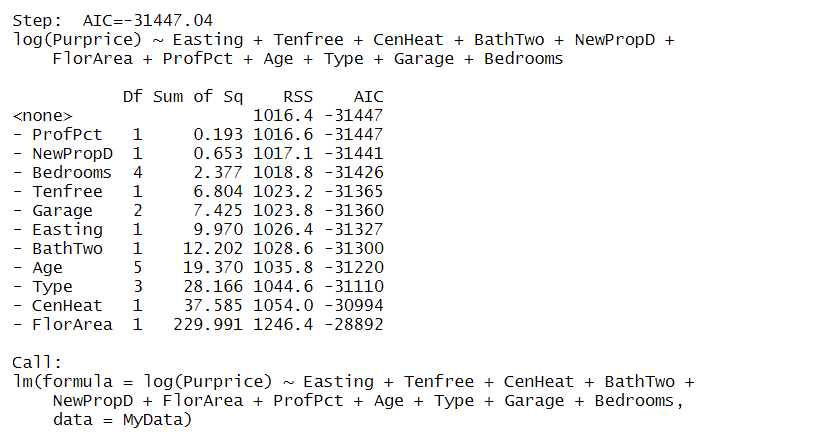
From the Correlation Matrix already plotted in the [Data preparation](#_Data_Preparation) section, Floor area is highly correlated to the purchase price. Other significant predictors are house type, age of the house and Lease hold/ Free hold indicators. Central heating and number of bedrooms are also equally determinant. Also Easting spatial variable has more significance than the Northing . Proportion of household with professional head (Profpct) and NewPropD are of least importance and can be ignored in the model predictions.

Below is the graph showing the variable importance using Random Forest Algorithm.



### Stepwise Regression:

Step wise procedure is used in backward direction for variable selection. From the summary statistics it is evident that FloorArea is most significant with the lowest AIC score. Variable Northing is removed from the model as it is not having influence over the house price. This is also witnessed by plotting the data to the map which is detailed under Prediction Modelling section below.



From the variable importance graph and stepwise Regression, dropping the variable “ProfPct” is of least significant and hence removed from the model. Similarly, variable “Northing” is not significant and removed from the stepwise regression results.

From the boxplots, residual plots Purchase price is skewed towards the right. Hence Log transformation applied on the response variable.

## Predictive Modelling

A very general methodology used in property pricing is knows as hedonic modelling. The property pricing is modelled as the function of

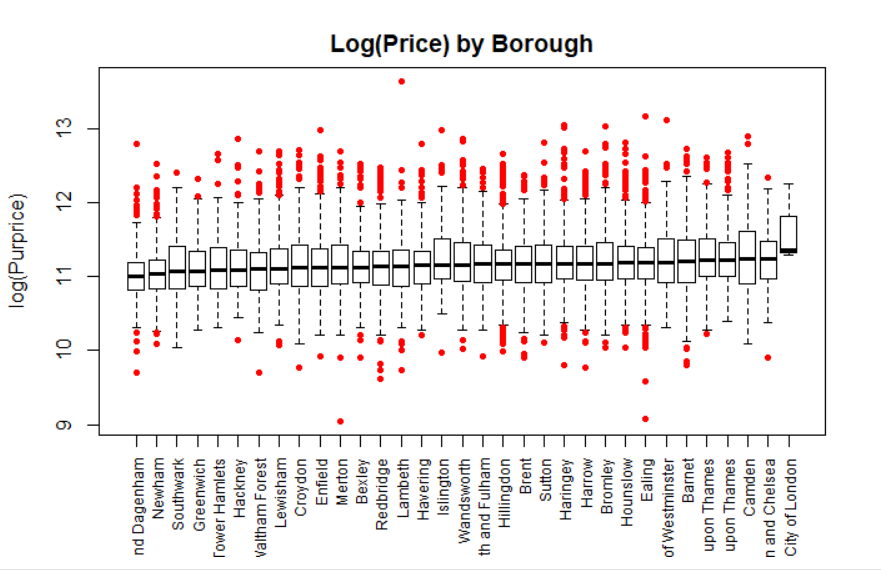
p=f(S,N,L)

Where p represents the property price, S represent the structural attributes represents a set of neighborhood attributes and L represent the location attributes. The interpretation of the parameter’s estimates plays an important role in the hedonic modelling. However hedonic modelling ignores the spatial effects in the house price modelling.

Here, the multilinear regression model was considering the response to be PurPrice which is the Purchase Price of the model with rest of the predictor. The AIC is used in the model to estimate the relative quality of the statistical model for a given dataset. Our main aim here is to find the predictor with the lowest AIC value. So, the model is fitted with all the variables included and the AIC value for each variable is checked. The model with least AIC value is considered to be the best fit model. The significance of the predictors of the best model provided by AIC is confirmed by checking their p-value. In this dataset, it is observed that the model with Purprice and Floor area gives the lowest AIC of 293198.7. Also, it can be confirmed by using p value. As per the scatterplot for AIC versus all the predictors, it can be brought to notice that New Property (NewPropD) variable has the Highest AIC and floor Area has the lowest AIC. Also we can check that the floor area seems to be very significant from the p value as it is very less than 0.05. Also, it can be noticed that the variables bedroom4, bedroom5, agebldpostw, agebld80s and profpct are the insignificant variables as they have a high p-value. Thus, all the other variables add significance to the model for the prediction of house prices.

When considering the role of the location of the property, we wanted to confirm whether the location of the houses play any significant roles in the prediction of prices of the house . This was performed by plotting the data to the map and it was inferred from the mapping that the houses of the London have comparatively low prices as we move towards the East of London. The housing prices are also slightly lower towards south of the city. But if move towards the west of the city, the house prices are comparatively increasing. Here, Easting variable of the house adds the most significance to the house price prediction. The interaction terms were also added to the model to see if the North-East or South-West add in any extra weightage or significance to the model, but as per the fit, we could see that the interaction terms does not add-in anything to the model. So, we can conclude that North-East facing is not significant for the modelling or prediction of the Property Pricing.

To explore the variation in the data, London borough’s data was considered and mapped with the point in polygon feature. The point in the map is considered to be the houses while the polygons are each borough of London. The point and polygon data were joined using spatial joins.The below plot is arranged in increasing order of the Log(Purprice) with each Borough of London.



As we know that logit function can provide a better visualization for the plots, logarithm of Purchase Price was considered and plotted with respect to Boroughs. From the plot, it can be recorded that the City of London has the highest Purchase price and the Dangenham has the lowest Purchase price among all the Boroughs . When looking at the log(Purchase Price) with respect to the types of houses and plotting it for better visualization, it can be noted that in the semi-detached type, Southwark has the highest Purchase price for the property and Dangneham Borough with the lowest Purchase price for the property. When looking at the flat type of houses, the average house price is the highest for the City of London, but there are no flats purchased in this area. The second highest Purchase price for the houses after the City of London Borough is for the Fulham and Camden Boroughs in London. It is also noticeable that many flats were purchased in these places and the least house price for the flat type is in Dagenham and Havering Boroughs in London.

**Standardized residuals**

To make sure that the prediction is accurate, the data was also checked for over-prediction and under-prediction. When looked at the boxplot between the standard residual and Boroughs of London, it appears that Hammersmith and Fulham, Kensington and Chelsea, Upon Thames and City of London house prices are over predicted and the Newham, Barking and Dagenham, Greenwich, Lewisham, Southwark, Havering, Redbridge, Croydon, Waltham Forest, Bexley and Bromley of London Boroughs , the house prices are under predicted. It could also be noticed from the histogram plot of count of houses with respect to the floor area that the most number of properties seems to be purchased which had a floor area of 700-800 meter square, the second most number of properties purchased had a floor area of 800-900 meter square and 900 - 950 meter square.

When the Borough median is plotted with respect to the standardized residuals, it can been observed from the plot that the Camden and Houslow Borough’s of London have the lowest standardized residual while Barking and Dagenham, Richmond, Lewisham and Greenwich Borough’s of London Residual have the highest rank among all the others. When the Floor Area is plotted with the Borough median, it can be inferred that the Boroughs namely Brent, Greenwich, Newham, City of Westminster are at their lowest while Haringey, Kingston, Islington and Hilington Boroughs are ranked among the highest. We can observe that all the residual values are higher than the median for each Borough. We have also conducted t-test on the standardized residuals , and noticed that the p value is greater than 0.05 , which indicates there is no normality between the residuals.

From discussion and analysis made above, these are the Findings from the above methods which can answer a few questions which may arise:

1) Since Easting and Northing plays a significant role in the prediction of property prices, it is very safe to conclude that the **Location** of the houses has a very evident impact on the house price prediction.

2) Residual: To check whether the residual of the fitted linear model differs from the beneficial properties of zero mean, independent, and homoscedastic, the residual diagnostic was performed over the linearly fitted model. From the plot, the following results can be clearly concluded:

**Residual vs Fitted Plot** : This is a scatter plot of residual (on y-axis) and fitted value (on x- axis) which is used to find any non-linearity, unequal residual variances (Heteroscedasticity) and outliers. From the residual vs Fitted plot, we could see that the mean appears to be zero which satisfies the first property of residual. Homoscedasticity is also fulfilled by this model. It can also be observed that there are few outliers namely Observation 4727, 5195 and 1415 as per the plot.

A close up of a map

Description automatically generated

**Normal Q-Q Plot** : This is a scatter plot created by plotting two different sets of quantiles against each other. The Q-Q plot helps us to asses if the data apparently comes from the Normal distribution. As per the plot, it can be noted that the curve deviates from the assumption of Normality. It also appears that the model does not follow linearity.

A close up of a map

Description automatically generated

**Residual vs Leverage Plot** : This plot helps to find out the influential points from the data. While outliers can be influential point, it is not necessary that they are. In other words, not all the outliers can influence the model largely while few points which are in the normal range can be highly influential. From the Residual vs Leverage Plot, it can be seen that there are few outliers such as 4727, 5195 and 1415 but nothing falls beyond the range of cook’s distance, so it does not influence the model’s prediction to any large extent.

A close up of a map

Description automatically generated

Durbin Watson Test is a test to check for the autocorrelation in the residual which in turn can tell us about the independence among the residuals. This test always provides a value between 0 to 4 where 2 means there is no correlation between residuals, less than 0 to 2 means that the positive autocorrelation exists and the values between 2 to 4 indicates negative correlation exists between the residuals. The value for Autocorrelation according to this test was observed as 0.0182 which shows that residuals are dependent. Also, the same can be confirmed from the p-value of test which is 0.028 which tells that there is a correlation existing between residual and hence, the residual are not independent of each other. Hence, the third property(independence) of the residual does not hold for this data.

lag Autocorrelation D-W Statistic p-value

1 0.01823648 1.963524 0.028

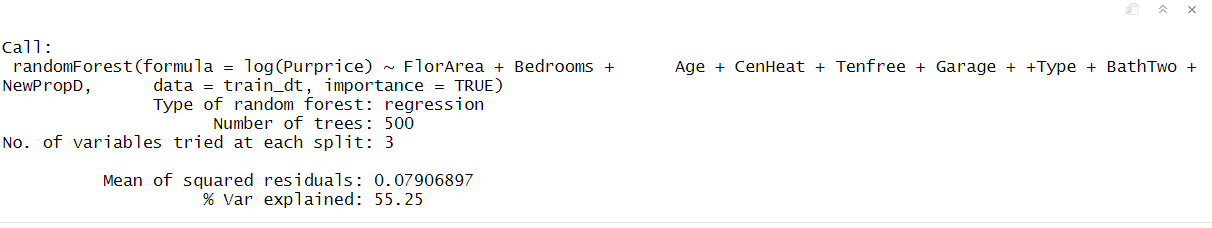
Alternative hypothesis: rho != 0

## Model Validation

Dataset is divided into training and test data set in the ratio of 60:40 to evaluate the train and test error. Reliability of the model is compared using the MSE (mean Squared errors), AIC scores for random forest, OLS and GWR.

#### **Random Forest**

Random Forest regression algorithm is used to fit the model. Variable importance is assessed and plotted to visualise the important determinants already discussed in the Descriptive analysis section. Attribute ProfPct is dropped from the models it has the least importance in the negative direction.



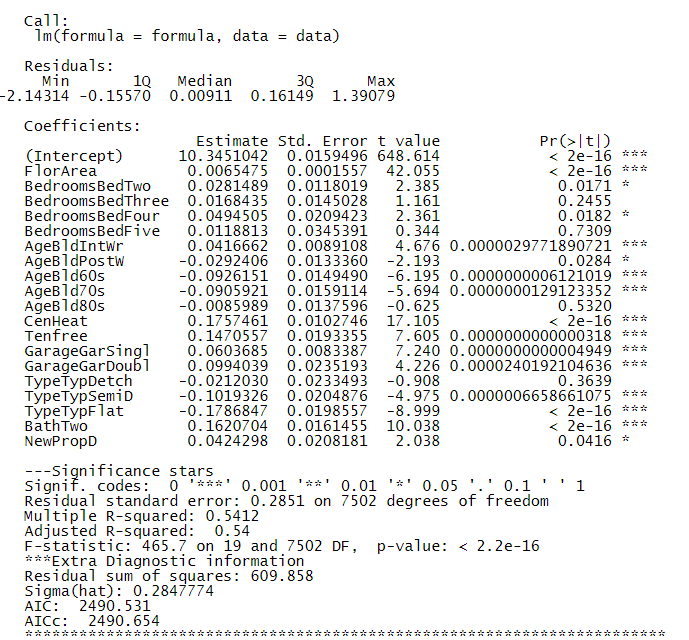
Train data is fitted to the below model and the train and test error are computed to check on goodness of fit and overfitting or under fitting issues.

#### **log(Purprice) = + FlorArea+Bedrooms +Age+CenHeat+Tenfree+ Garage + Type + BathTwo + NewPropD +Error€**

Where **0**is the Intercept the base price of the house when all the other predictors are 0. ****are the Co-efficient of the predictor variable that explains the increase in Purprice for a unit increase in predictors. This model is having the test error 0.077 and train error of 0.053. Same model is fitted using Ordinary Least square below.

#### **Ordinary Least Squares**

Multiple linear regression is fitted to the model as there are multiple predictors to determine the purchase price of the housing. The regression coefficients are assessed for significance again on the summary statistics.



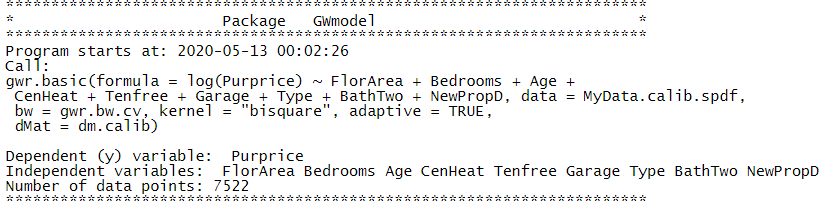
MLR is having the train error of 0.083 and test error of 0.080

#### **Geographically Weighted Regression**

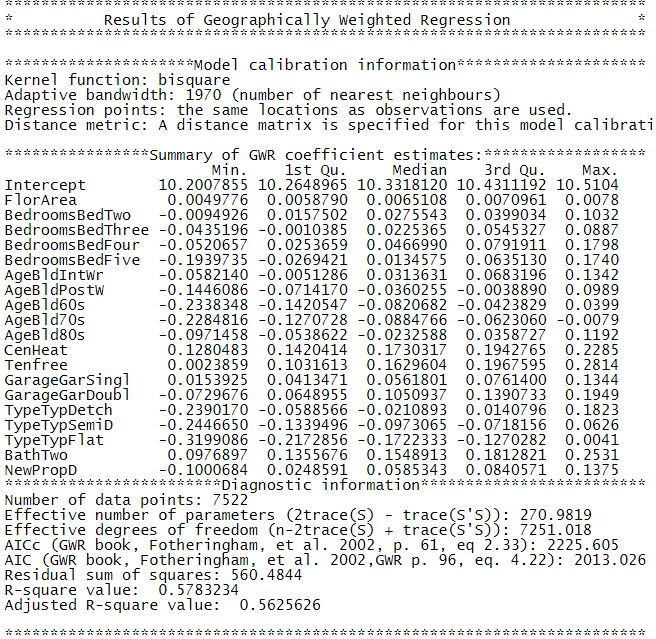
Using Geographically weighted models we make estimates of the parameters at each Location. The London data frame is converted to a spatial points dataframe with Easting and Northing attributes as co-ordinates and mapped to the London borough regions using the London shape files.

Train and test data (Sampling in 60:40 ratio) are passed to geographically weighted regression. This model is specifically used to predict any spatial relation and compares the results with ordinary Least Squares.

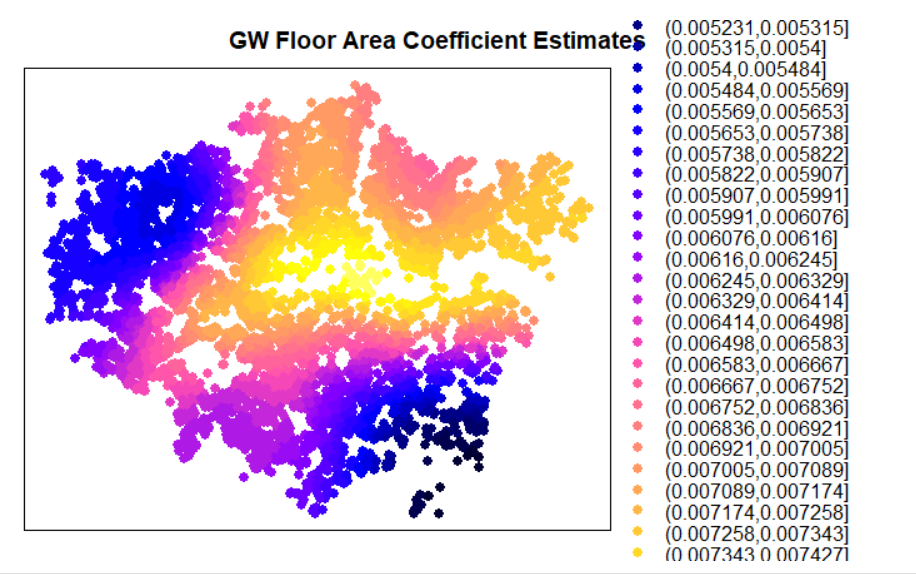
GW Model is fitted with same regression equation used for Random forest and OLS.



**Summary Statistics of GWR:**



The coefficient estimates plot of most determinant predictor Floor Area using GW regression reveals that house prices are high when we move towards the centre of London and lowest at the south east and north west regions. Comparing the AIC and R-squared, Adjusted R-squared value of all three models GW is having the optimal results with lowest AIC and explained variance. This model can be further explored to have interaction terms and compared for better predictions.



## Conclusions

From the analysis, the top5 determinants that influence the purchase price are Floor Area, number of bedrooms, age of the building, central Heating and Leasehold Indicator. There are other attributes like type of the house, number of garages , number of baths, contributes to the price prediction but not highly significant. Also insignificant variables were removed and the model was created.Purchase Price is highly correlated to Floor Area and accounts for almost 70% of the price prediction in the London dataset. GWR adds the most weightage in the prediction of the housing prices. Exploratory study of spatial correlation with purchase price using GWR and plotting the predictions reveals that the houses of the London have comparatively low prices as we move towards the East of London. The housing prices are also slightly lower towards south of the city. But if move towards the west of the city, the house prices are comparatively increasing. Here, Easting variable of the house adds the most significance to the house price prediction. Also, the Random forest provided a better prediction as compared to OLS in Multilinear Regression as the test error is lower in prediction using Random Forest.