NCG612 - Housing Project

Housing Valuation: finding the most reliable determinations on property price for Greater London

Group H

Haojun He (19250816)

An Ning Shen

Yang Wang (19250003)

Contents

1 Introduction

Housing Valuation is an area in which statistical models can play a role. The models which are frequently used can also be used to model other price structures. The project is concerned with finding the most reliable determinants of property prices. The dataset is a subset of anonymized mortgage records for the area that is known as Greater London. The purchase price (which is different from the asking price) is available, as a series of characteristics of the property. The goal is to find the best group of predictors of property prices and to find the most reliable determinants of property prices.

Describing the methods for property price prediction

- Obtain significant predictor variables in predicting prices of housing in London
- Obtain an estimate of the spatial variation in the influence of floorspace change on the price by borough.
- Geographically weighted regression (GWR) is a unique type of regression. Compared to a linear regression, the predictors contribute a coefficient value which tells how much the response is changed based on a unit change in the predicting variable. Whereas in GWR, the coefficient value changes based on spatial orientation. A coefficient value is no longer global and is calculated based on that specific region. This will decrease bias and give out a more intriguing and accurate response and analysis,

2 Data Preparation

2.1 Variables in Original Dataset

	Name	Type	Description
1	X	int	No.
2	Easting	int	Easting in m
3	Northing	int	Northing in m
4	Purprice	int	Purchase Price in GBP
5	BldIntWr	int	Built between 1918 and 1939
6	BldPostW	int	Built between 1945 and 1959
7	Bld60s	int	Built between 1960 and 1969
8	Bld70s	int	Built between 1970 and 1979
9	Bld80s	int	Built between 1980 and 1989
10	TypDetch	int	Detached property
11	TypSemiD	int	Semi-detached property
12	TypFlat	int	Flat or apartment
13	GarSingl	int	Single Garage
14	GarDoubl	int	Double Garage
15	Tenfree	int	Leasehold/Freehold indicator
16	CenHeat	int	Central heating
17	BathTwo	int	Two or more bathrooms
18	BedTwo	int	Two bedrooms
19	BedThree	int	Three bedrooms
20	BedFour	int	Four bedrooms
21	BedFive	int	Fie bedrooms
22	NewPropD	int	New property
23	FlorArea	double	Floor area in square metres
24	NoCarHh	double	Proportion of households without a car
25	CarspP	double	Cars per person in neighborhood
26	ProfPct	double	Proportion of Households with Professional Head
27	UnskPct	double	Proportion of Households with Unskilled head
28	RetiPct	double	Proportion of residents retired
29	Saleunem	double	Not known
30	Unemploy	double	Unemployed workers
31	PopnDnsy	double	Local population density

2.2 Data Cleaning

Convert dummies to factors - more convenient for modelling.

For building a model to predict the price of a property in London, some variables should be organized properly.

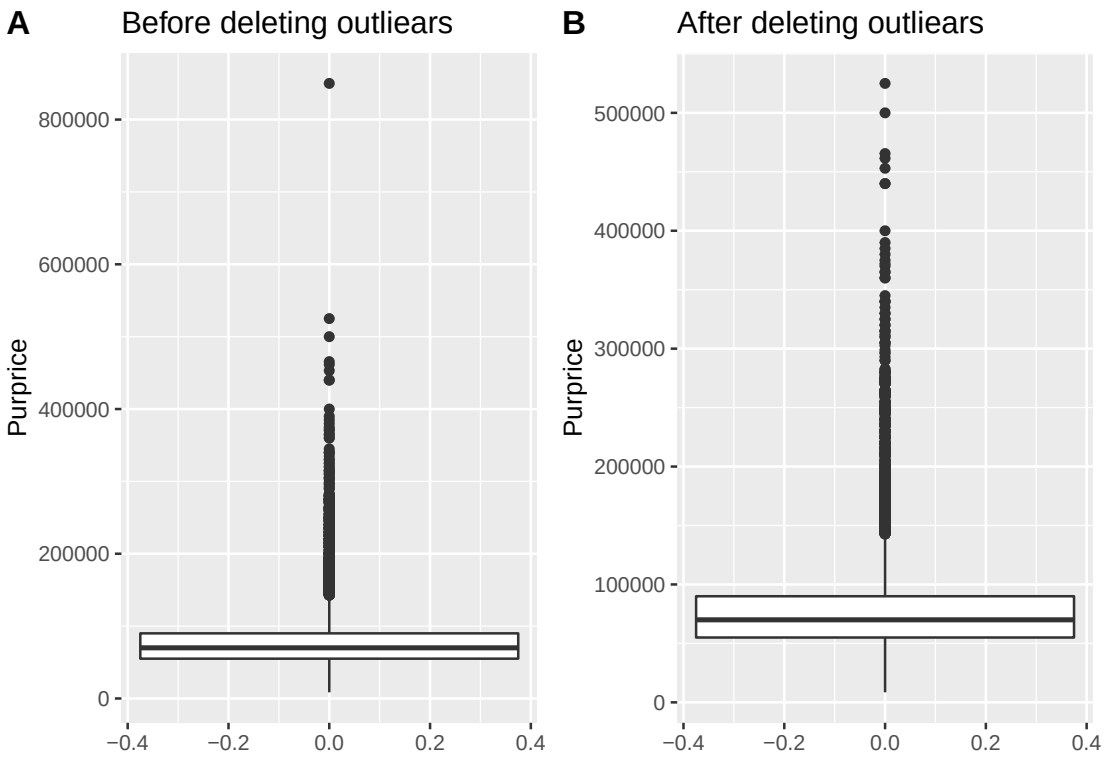
- Age: these represent the time period in which the property was constructed. It is from variables BldIntWr,BldPostW,Bld60s,Bld70s and Bld80s. The values of it are PreWW1, BldIntWr, BldPostW, Bld60s, Bld70s and Bld80s.
- Type: these represent the type of building. It is from variables TypDetch,TypSemiD and TypFlat. The values of it are TypDetch, TypSemiD, TypFlat and Bungalow.
- Garage: these represent the numbers of garage that the property has. It is from variables GarSingl and GarDoubl. The values of it are HardStnd, GarSingl and GarDoubl.
- Bedrooms: these represent the numbers of bedrooms that the property has. It is from variables BedTwo, BedThree, BedFour and BedFive. The values of it are BedOne, BedTwo, BedThree, BedFour and BedFive.

2.3 Variables in New Dataset.

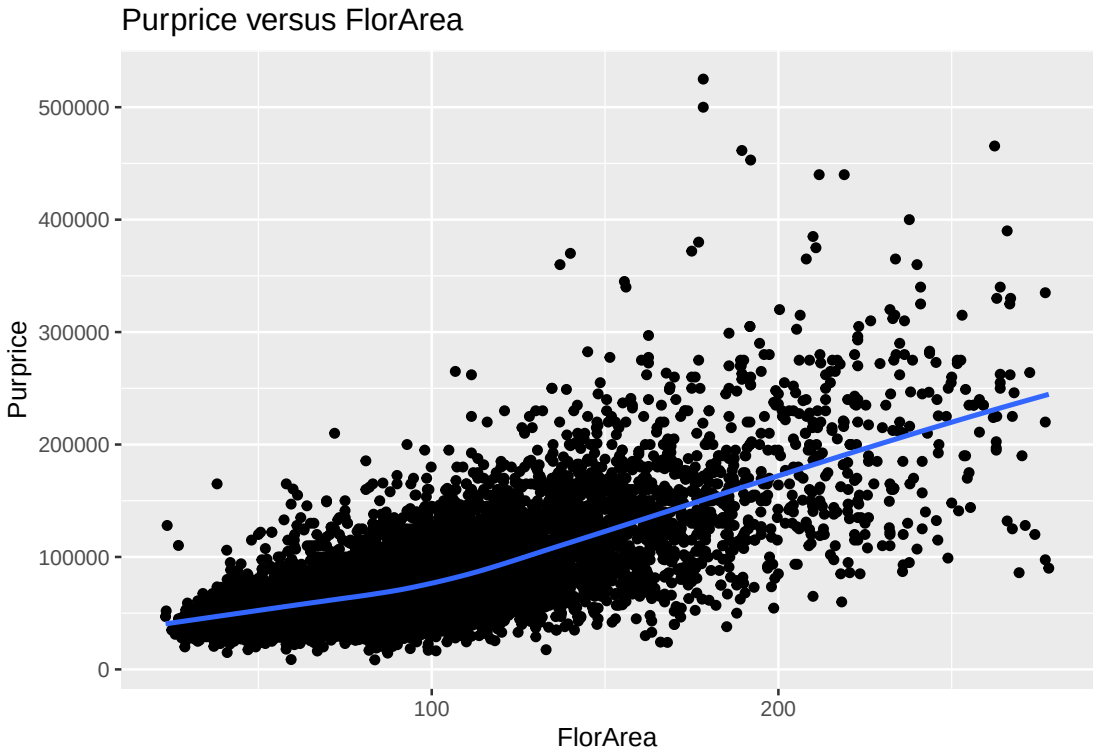
```
##        Name   Type                                        Description
## 1   Easting    int                                       Easting in m
## 2  Northing    int                                      Northing in m
## 3  Purprice    int                                Purchase Price in GBP
## 4   Tenfree    int                       Leasehold/Freehold indicator
## 5   CenHeat factor                                    Central heating
## 6   BathTwo factor                                Two or more bathrooms
## 7  NewPropD    int                                        New property
## 8  FlorArea double                          Floor area in square metres
## 9   ProfPct double Proportion of Households with Professional Head
## 10      Age factor                                  The age of building
## 11     Type factor                                 The type of building
## 12   Garage factor                               The Garage of building
## 13 Bedrooms factor                               The number of Bedrooms
## 14  NoCarHh double             Proportion of households without a car
## 15   CarspP double                    Cars per person in neighborhood
## 16  UnskPct double    Proportion of Households with Unskilled head
## 17  RetiPct double                    Proportion of residents retired
## 18 Saleunem double                                           Not known
## 19 Unemploy double                                 Unemployed workers
## 20 PopnDnsy double                           Local population density
```

3 Data Exploration

3.1 Exploration of Response Variable (Purprice)

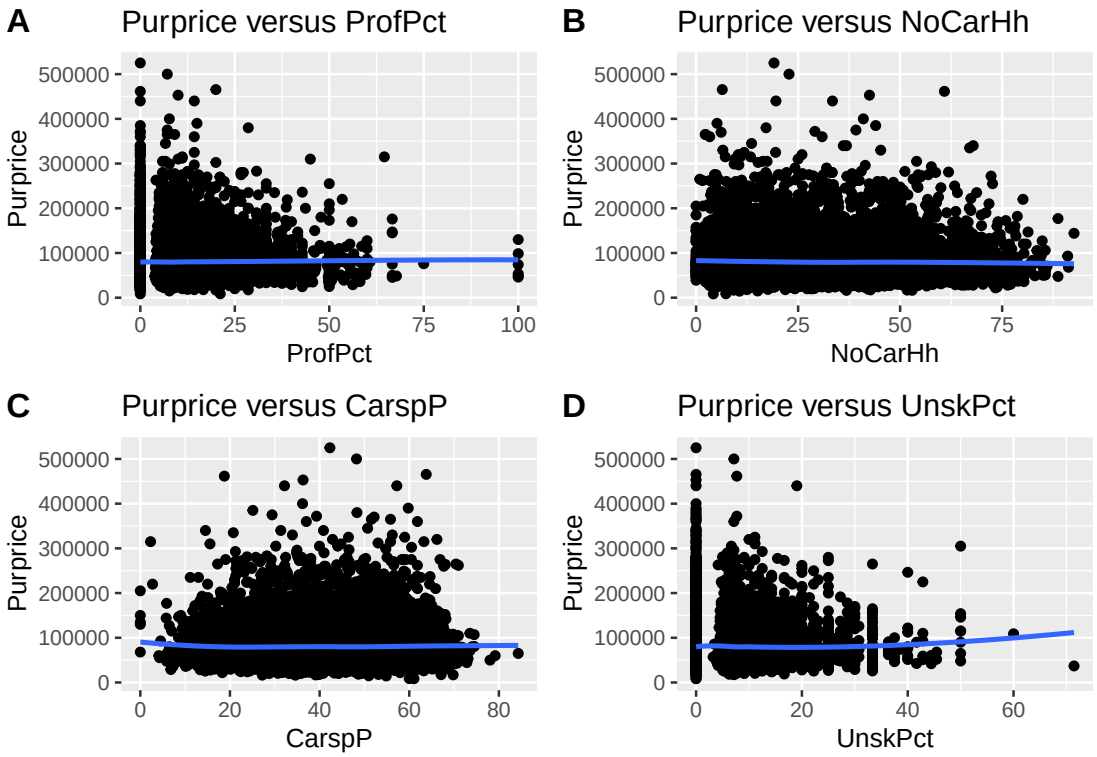


Delete the outlier which is over 600,000. Most of the prices of the property are under 600,000, but there is an outlier, which is much bigger than others. It would influence the result of the analysis.

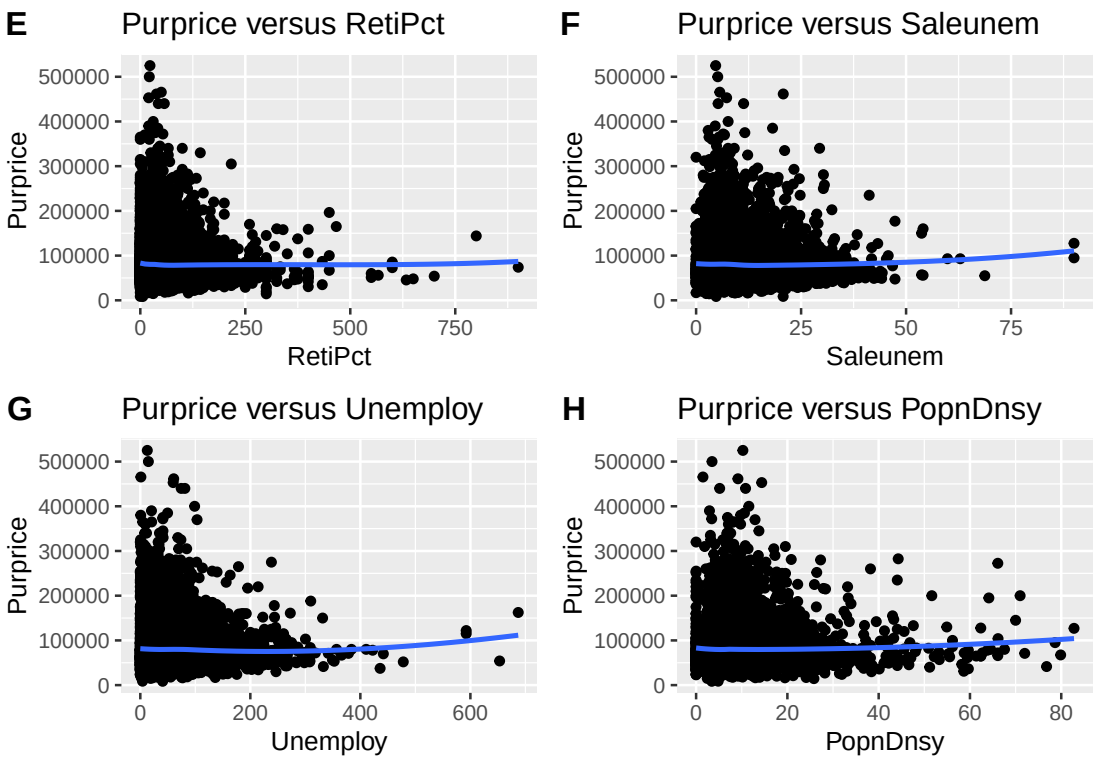


The floor Area and price show a somewhat linear relationship. The slope is constant and no clear curvature is present. The price increases as the floor area increases.

3.2 Exploration of Independent Variable (Continous Variables)



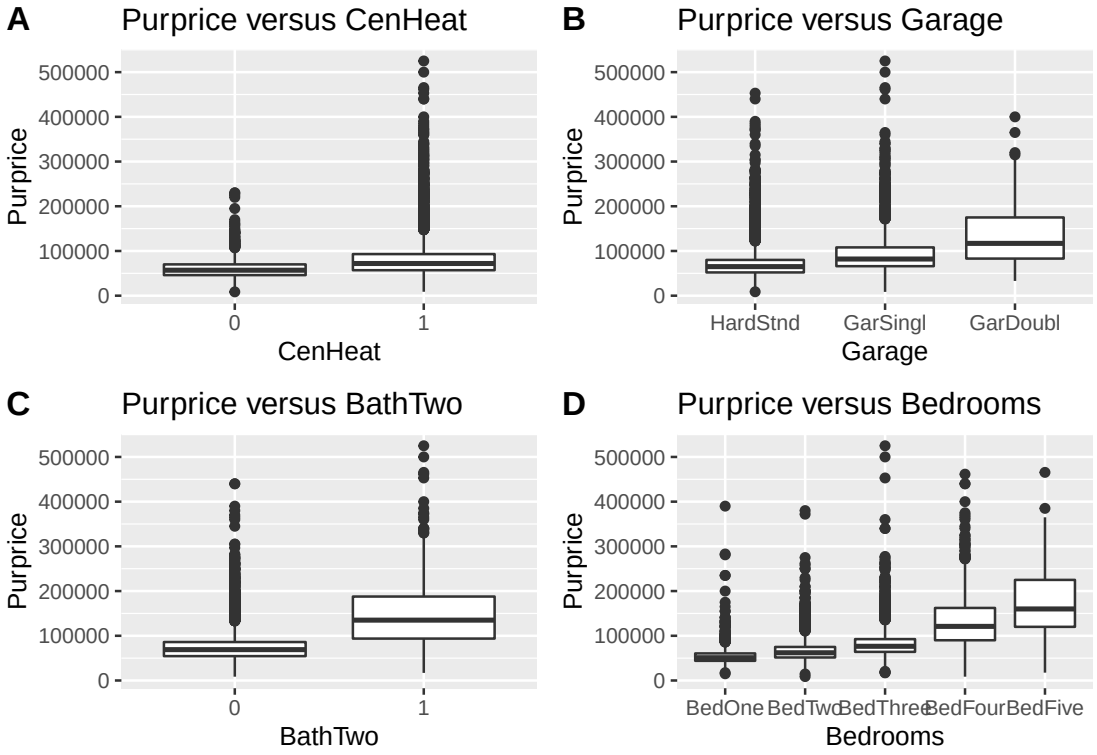
- A: For Profpct, only a few observations have higher values and a linear relationship is inadequate.
- B: There is no outlier in NoCarHh, and no linear relationship between NoCarHh and Purprice.
- C: The line is almost horizontal, no linear relationship between CarspP and Purprice.
- D: Most observations have lower values in UnskPct, the trend of the line is influenced by outliers. The linear relationship is inadequate.



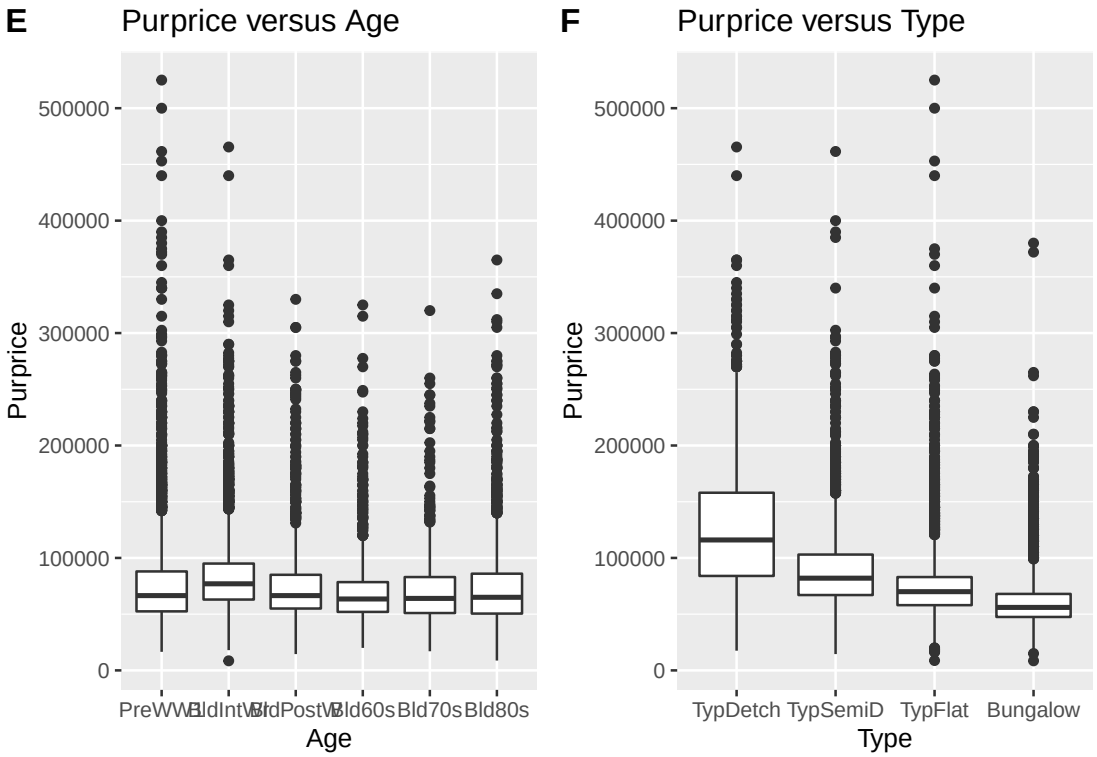
- E: We can see in the plot, the line is horizontal. It means there is no linear relationship between RetiPct and Purprice.
- F: The trend of the line in this plot is influenced by a few points. The linear relationship is inadequate.
- G: The trend of the line in this plot is influenced by a few points. The linear relationship between Unemploy and Purprice.
- H: The plot shows the relationship between PopnDnsy and Purprice is very weak.

Based on the plots above, all these variables do not have a strong linear relationship that dependent variable property price. For cars per person in neighborhood and proportion of households with unskilled heads, the fit line almost horizontal. It means They have no liner relationship with property price. The other variables are scattered around the origin, and most of the points are scattered tightly around the origin. The trend of lines is mainly influenced by outliers. In the same conclusion we can get that the relationship between them and the price of the property is very weak.

3.3 Exploration of Independent Variable (Categorical Variables)



- A: It shows that houses with central heating are higher priced than houses without central heating. Although the average price of houses with central heating is higher, it does not differ by a large price difference. It is more comfortable when heating is provided 24/7 as to heating which needs to be set up before using which could cause discomfort in some cases.
- B: From the number of garages, we can clearly see that the houses with two garage's median prices are a lot higher than houses with a single garage. Again, the size of the house is influenced by how many cars the garage can park. By assumption, one wouldn't have two garages with a single room. It would only be available to houses with more than two rooms to have two garages.
- C: Furthermore, we can see that houses with two bathrooms are also higher priced on average. This difference between one bathroom to two bathrooms is much higher. Intuitively, this would be more convenient and houses with more than one washroom are typically bigger in size based on the design of the interior.
- D: Finally moving on to the number of bedrooms a house would have. We can see that the houses with one room and two rooms do not differ by much. Even three rooms do not have too much difference in the median of pricing. However, as the bedroom goes to four or even five, the jump is significantly higher.



- E: Moving on to the next predictor, we have the age of the house. From our plots, we can see that housing before World War 1 has the greatest span of pricing. It is usually because the location of the housing was excellent since it was just the beginning. Therefore, it could be one of the reasons to explain the span of prices.‘
- F: The type of the house also influences the pricing of housing. For example, we had detached homes, semidetached and flats. Obviously detached homes would have the highest pricing, as it has more privacy and the layout of the houses are better. Then we have the semidetached, which is still good. However, it does lack the same amount of privacy from a fully detached house. Flats would be at the end of the list since there is little privacy if the isolation was not done well.

In the plots above, we can see that the types of property are an important factor that influences the price of a property. The property with central heating tends to be more expensive. As the number of garages, bathrooms and bedrooms goes up, the price of a property shows an increasing trend. However, the age of the property seems to have not to influence on the price of the property. The large houses clearly cost more, however as the size of the houses goes up, there are few data available. As we can see from our PurPrice vs FloorArea plot, the left size is tightly scattered with data and the right side of the line has a lot fewer data.

4 Fit Linear Models

4.1 Introduction of Model

With all the predictors examined, we move to our simple linear regression model.

We first use lm() function in R for our models.

- lm(Purprice~., data=MyData)

If we were to write out the function, it would be :

- Purprice = b0 + b1*FlorArea + b2*Bedrroms + b3*Type +... + b17*PopnDnsy

Our predictors would be able to predict the price of a house based on given London data. It would be able to predict the price based on the coefficients of the predictors. It is only required to have the right input to predict the price.

Then we want to find the predictor that has the most impact on price. So, we used AIC to compare the different predictors. Then fit model with all predictors and choose significant predictors for the linear model.

Finally, fit model with significant predictors and check VIF of predictors to avoid colinearity.

4.2 Fit for A Single Variable and Look at AICs

```
##        name      AIC
## 5  FlorArea 293198.7
## 10 Bedrooms 297086.6
## 8      Type 299031.9
## 3   BathTwo 299667.9
## 9    Garage 300656.4
## 1   Tenfree 300712.3
## 2   CenHeat 301562.2
## 7       Age 301833.0
## 14  RetiPct 301924.9
## 11  NoCarHh 301925.5
## 16 Unemploy 301928.1
## 15 Saleunem 301928.6
## 6   ProfPct 301929.4
## 12   CarspP 301929.6
## 17 PopnDnsy 301930.1
## 4  NewPropD 301930.4
## 13  UnskPct 301931.5
```

To choose significant variables for the model, we build a model for response and every predictor respectively and output the AICs of models in the table above. We can see that the area of the floor is the most important predictor for predicting the price of properties. The number of bedrooms, bathrooms and property types are also impacted the property price greatly.

4.3 Fit Linear Model With All Predictors

```
##
## Call:
## lm(formula = Purprice ~ ., data = MyData[, 3:20])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -136420  -13483   -1322   10340  371624
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)       12216.791   3234.459   3.777             0.000159 ***
## Tenfree1           6215.453   1352.354   4.596    0.000004348433111 ***
## CenHeat1          11856.974    754.564  15.714 < 0.0000000000000002 ***
## BathTwo1          24029.638   1202.984  19.975 < 0.0000000000000002 ***
## NewPropD1          1879.988   1545.371   1.217             0.223807
## FlorArea            677.818     11.379  59.570 < 0.0000000000000002 ***
## ProfPct              45.261     24.909   1.817             0.069235 .
## AgeBldIntWr        3996.542    657.059   6.082    0.000000001217999 ***
## AgeBldPostW       -1134.041    975.441  -1.163             0.245017
## AgeBld60s         -7318.654   1089.956  -6.715    0.000000000019669 ***
## AgeBld70s         -6713.012   1164.455  -5.765    0.000000008361791 ***
## AgeBld80s           357.120   1026.580   0.348             0.727941
## TypeTypSemiD    -12406.759   1005.914 -12.334 < 0.0000000000000002 ***
## TypeTypFlat     -17328.944   1032.138 -16.789 < 0.0000000000000002 ***
## TypeBungalow     -5754.224   1658.808  -3.469             0.000524 ***
## GarageGarSingl    3773.963    614.680   6.140    0.000000000851699 ***
## GarageGarDoubl    9279.791   1676.432   5.535    0.000000031670020 ***
## BedroomsBedTwo   -3399.740    869.034  -3.912    0.000091984407889 ***
## BedroomsBedThree -7863.395   1068.092  -7.362    0.000000000000192 ***
## BedroomsBedFour  -1709.352   1542.268  -1.108             0.267738
## BedroomsBedFive   3973.657   2504.121   1.587             0.112573
## NoCarHh            -12.783     30.913  -0.414             0.679247
## CarspP             -19.105     40.555  -0.471             0.637592
## UnskPct            -40.730     36.679  -1.110             0.266830
## RetiPct             -7.091      5.618  -1.262             0.206928
## Saleunem            54.662     60.992   0.896             0.370154
## Unemploy             9.737      5.808   1.677             0.093629 .
## PopnDnsy            42.660     40.423   1.055             0.291287
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27150 on 12507 degrees of freedom
## Multiple R-squared:  0.5652, Adjusted R-squared:  0.5643
## F-statistic: 602.2 on 27 and 12507 DF,  p-value: < 0.00000000000000022
```

Then fit the linear model with all predictors. The output of the model shows the proportion of households without a car, cars per person in the neighborhood, the proportion of households with professional head, the proportion of households with unskilled head, the proportion of residents retired, unemployed workers, the new properties and local population density are not significant. This conclusion the same as what we get in the correlation coefficient table. So these variables are moved out of the model.

4.4 Fit Linear Model With Significant Predictors and Check VIF

```
##
## Call:
## lm(formula = Purprice ~ Tenfree + CenHeat + BathTwo + FlorArea +
##     Age + Type + Garage + Bedrooms, data = MyData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -135607  -13414   -1328   10382  371167
##
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
## (Intercept)       12225.50    2113.53   5.784    0.000000007451054 ***
## Tenfree1           6140.30    1352.02   4.542    0.000005635720994 ***
## CenHeat1          11854.63     754.65  15.709 < 0.0000000000000002 ***
## BathTwo1          24053.12    1202.55  20.002 < 0.0000000000000002 ***
## FlorArea            678.08      11.37  59.613 < 0.0000000000000002 ***
## AgeBldIntWr        4051.07     656.87   6.167    0.000000000716479 ***
## AgeBldPostW       -1136.03     975.15  -1.165             0.244050
## AgeBld60s         -7336.41    1089.76  -6.732    0.000000000017450 ***
## AgeBld70s         -6707.95    1164.43  -5.761    0.000000008572948 ***
## AgeBld80s           935.35     898.71   1.041             0.298003
## TypeTypSemiD     -12381.18    1005.73 -12.311 < 0.0000000000000002 ***
## TypeTypFlat      -17269.10    1031.65 -16.739 < 0.0000000000000002 ***
## TypeBungalow      -5695.79    1658.13  -3.435             0.000594 ***
## GarageGarSingl     3776.80     614.47   6.146    0.000000000816264 ***
## GarageGarDoubl     9257.72    1676.29   5.523    0.000000034042618 ***
## BedroomsBedTwo    -3450.07     869.10  -3.970    0.000072362613187 ***
## BedroomsBedThree  -7869.64    1068.21  -7.367    0.000000000000185 ***
## BedroomsBedFour   -1743.82    1541.90  -1.131             0.258095
## BedroomsBedFive    3937.18    2504.29   1.572             0.115935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27150 on 12516 degrees of freedom
## Multiple R-squared:  0.5647, Adjusted R-squared:  0.564
## F-statistic: 901.9 on 18 and 12516 DF,  p-value: < 0.00000000000000022
```

```
##  Tenfree  CenHeat  BathTwo FlorArea      Age     Type   Garage Bedrooms
##    6.722    1.030    1.253    3.032    1.561    9.769    1.480    4.112
```

Buiding model with all significant predictors and check colinearity by VIF. In the table above, the colinearity of property type is very high(9.769). It should be moved out of the model. In the next step, the dataset would be separated into training and testing data and the linear model would be built using a training dataset and be tested using a testing dataset.

4.5 Fit Linear Model and Test Accuracy

```
##
## Call:
## lm(formula = Purprice ~ FlorArea + Bedrooms + BathTwo + Garage +
##     Tenfree + CenHeat + Age, data = trainData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -131677  -13606   -1649   10507  364562
##
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        2386.22    1529.39   1.560             0.118744
## FlorArea            715.26      14.97  47.766 < 0.0000000000000002 ***
## BedroomsBedTwo    -5323.44    1148.02  -4.637   0.0000035935589099 ***
## BedroomsBedThree -10485.73    1398.61  -7.497   0.0000000000000726 ***
## BedroomsBedFour   -2291.56    2049.23  -1.118             0.263493
## BedroomsBedFive    5954.67    3305.12   1.802             0.071640 .
## BathTwo1          24198.54    1585.61  15.261 < 0.0000000000000002 ***
## GarageGarSingl     6137.83     793.09   7.739   0.0000000000000113 ***
## GarageGarDoubl    15060.80    2180.87   6.906   0.0000000000053983 ***
## Tenfree1          -1618.72     939.73  -1.723             0.085013 .
## CenHeat1          12268.51    1000.50  12.262 < 0.0000000000000002 ***
## AgeBldIntWr        5663.90     858.19   6.600   0.0000000000439682 ***
## AgeBldPostW        2312.66    1279.79   1.807             0.070791 .
## AgeBld60s         -5397.07    1436.10  -3.758             0.000172 ***
## AgeBld70s         -5712.37    1522.99  -3.751             0.000178 ***
## AgeBld80s          3592.95    1174.30   3.060             0.002224 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27920 on 7505 degrees of freedom
## Multiple R-squared:  0.5553, Adjusted R-squared:  0.5544
## F-statistic: 624.8 on 15 and 7505 DF,  p-value: < 0.00000000000000022

## [1] 777827779

## [1] 723418875
```

As the output of the model above, the mean square error of the testing dataset is 777827779 which is slightly lower than that of the training dataset. For predictor floor area, 1 square metre increase, the average price of the property would increase 715.26 GBP, keeping other predictors constant. The average price for those properties with central heating is higher than those without central heating by 12268.51 GBP, keeping other predictors constant.

5 Spatial Variation

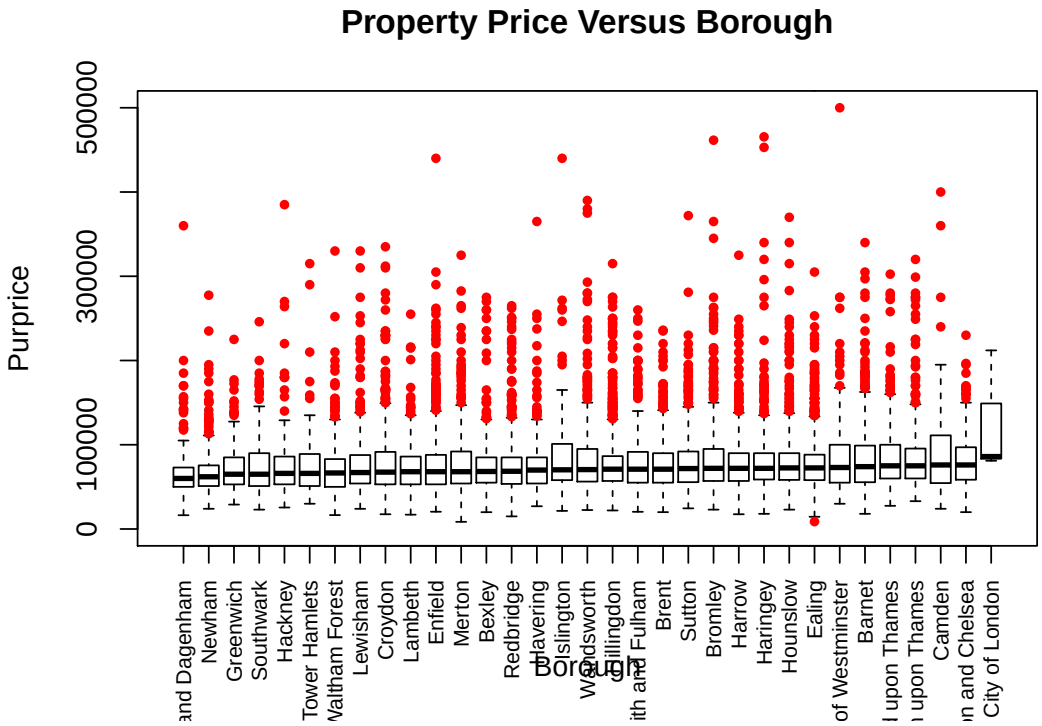
5.1 Fit Model with Variable Easting and Northing

```
##
## Call:
## lm(formula = Purprice ~ x + y + I(x^2) + I(y^2) + I(x * y), data = MyData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73924 -24782  -9828   9862 444261
##
## Coefficients:
##                  Estimate   Std. Error t value  Pr(>|t|)
## (Intercept) -3153110.597   874075.711  -3.607  0.000311 ***
## x              12253.587     2793.148   4.387 0.0000116 ***
## y                352.539     2915.762   0.121  0.903766
## I(x^2)           -10.739        2.555  -4.203 0.0000266 ***
## I(y^2)             7.372        4.717   1.563  0.118080
## I(x * y)          -5.727        4.323  -1.325  0.185350
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41050 on 12529 degrees of freedom
## Multiple R-squared:  0.004172,   Adjusted R-squared:  0.003774
## F-statistic:  10.5 on 5 and 12529 DF,  p-value: 0.0000000004507
```

Fitting model with variable Easting and Westing to test is the location influencing the price of properties significantly. The result shows that the properties tend to have a lower price as we move east and the influence is significant. So it is necessary to consider the geographic effect in predicting the price of properties.

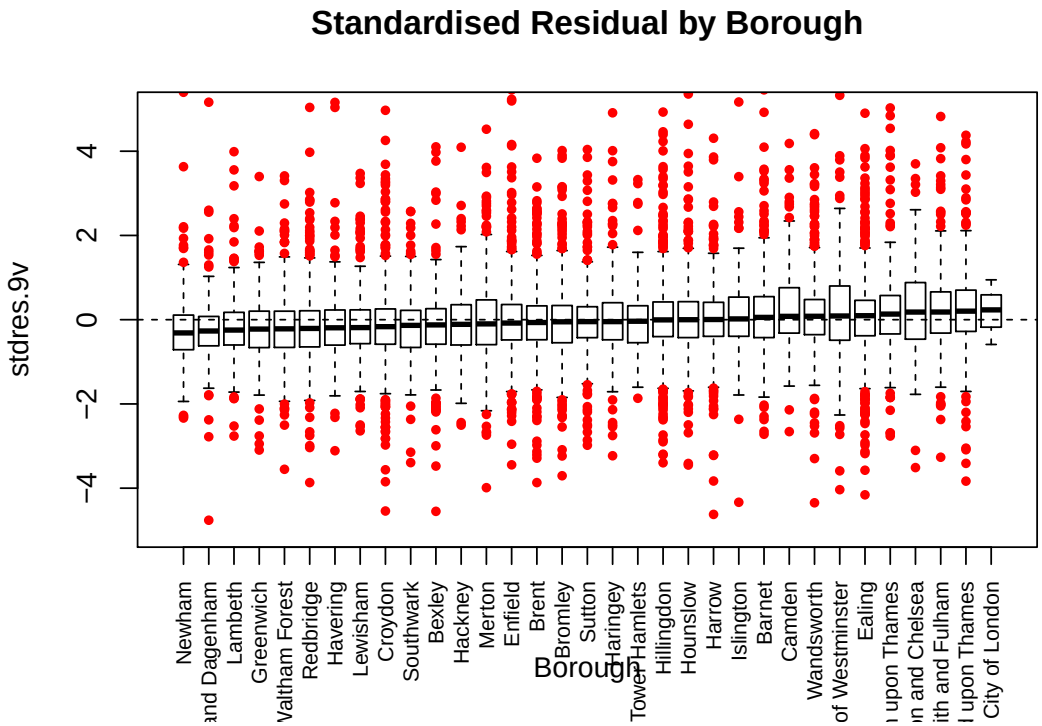
5.2 Load Borough Data

5.2.1 Property Price Versus Borough



In the plot above, We can see the median property price is different in different boroughs in London. Especially in the city of London, property price is significantly higher than that in other boroughs.

5.2.2 Standardised Residuals Versus Borough



In the borough versus standard residual plot, we can get the same conclusion that the distribution of residuals in different boroughs are different. If we can fit model considering the effect from boroughs, the result might be better.we will now run a geographically weighted regression model to see how the coefficients of the model might vary across London.

5.3 Geographically Weighted Regression (GWR)

5.3.1 Fit GWR Model

First we will calibrate the bandwidth of the kernel that will be used to capture the points for each regression (this may take a little while) and then run the model:

```
##    *********************************************************************
##    *                       Package   GWmodel                            *
##    *********************************************************************
##    Program starts at: 2020-05-14 10:14:26
##    Call:
##    gwr.basic(formula = Purprice ~ FlorArea + Bedrooms + BathTwo +
##     Garage + Tenfree + CenHeat + Age, data = map, bw = bw, kernel = "gaussian")
##
##    Dependent (y) variable:  Purprice
##    Independent variables:  FlorArea Bedrooms BathTwo Garage Tenfree CenHeat Age
##    Number of data points: 7521
##    *********************************************************************
##    *                    Results of Global Regression                   *
##    *********************************************************************
##
##    Call:
##     lm(formula = formula, data = data)
##
##    Residuals:
##     Min      1Q  Median      3Q     Max
## -131677  -13606   -1649   10507  364562
##
##    Coefficients:
##                      Estimate Std. Error t value             Pr(>|t|)
##    (Intercept)        2386.22    1529.39   1.560             0.118744
##    FlorArea            715.26      14.97  47.766 < 0.0000000000000002 ***
##    BedroomsBedTwo    -5323.44    1148.02  -4.637   0.0000035935589099 ***
##    BedroomsBedThree -10485.73    1398.61  -7.497   0.0000000000000726 ***
##    BedroomsBedFour   -2291.56    2049.23  -1.118             0.263493
##    BedroomsBedFive    5954.67    3305.12   1.802             0.071640 .
##    BathTwo1          24198.54    1585.61  15.261 < 0.0000000000000002 ***
##    GarageGarSingl     6137.83     793.09   7.739   0.0000000000000113 ***
##    GarageGarDoubl    15060.80    2180.87   6.906   0.0000000000053983 ***
##    Tenfree1          -1618.72     939.73  -1.723             0.085013 .
##    CenHeat1          12268.51    1000.50  12.262 < 0.0000000000000002 ***
##    AgeBldIntWr        5663.90     858.19   6.600   0.0000000000439682 ***
##    AgeBldPostW        2312.66    1279.79   1.807             0.070791 .
##    AgeBld60s         -5397.07    1436.10  -3.758             0.000172 ***
##    AgeBld70s         -5712.37    1522.99  -3.751             0.000178 ***
##    AgeBld80s          3592.95    1174.30   3.060             0.002224 **
##
##    ---Significance stars
##    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##    Residual standard error: 27920 on 7505 degrees of freedom
##    Multiple R-squared: 0.5553
##    Adjusted R-squared: 0.5544
##    F-statistic: 624.8 on 15 and 7505 DF,  p-value: < 0.00000000000000022
```

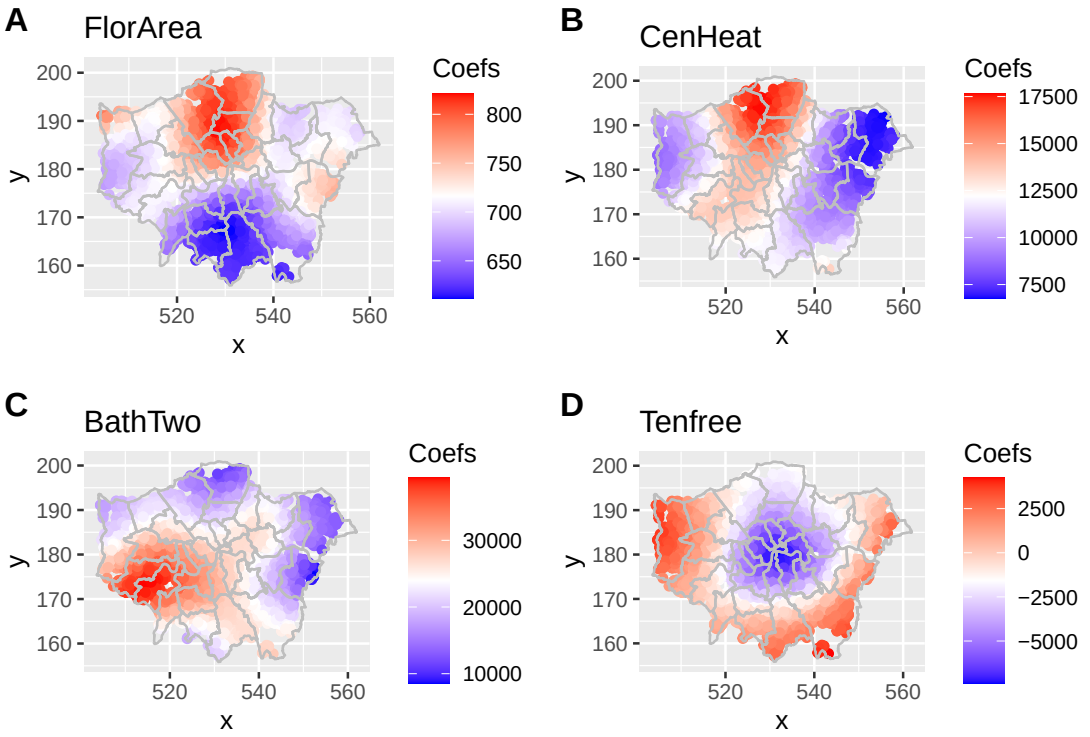
```
##    ***Extra Diagnostic information
##    Residual sum of squares: 5850042725710
##    Sigma(hat): 27893.27
##    AIC:  175347.7
##    AICc:  175347.8
##    ***********************************************************************
##    *          Results of Geographically Weighted Regression              *
##    ***********************************************************************
##
##    *********************Model calibration information*********************
##    Kernel function: gaussian
##    Fixed bandwidth: 6103.785
##    Regression points: the same locations as observations are used.
##    Distance metric: Euclidean distance metric is used.
##
##    ****************Summary of GWR coefficient estimates:******************
##                            Min.    1st Qu.     Median    3rd Qu.     Max.
##    Intercept        -11554.796    254.395   4904.294   7410.595 11034.18
##    FlorArea            612.043    676.780    707.772    732.984   821.42
##    BedroomsBedTwo   -12787.458  -6204.371  -4945.664  -3734.715 -1174.21
##    BedroomsBedThree -20027.410 -12955.621 -10648.990  -7608.165 -3183.76
##    BedroomsBedFour  -16583.906  -7625.159  -3919.857   3315.564  9841.25
##    BedroomsBedFive  -27505.494 -11882.768   8012.577  16527.424 80111.69
##    BathTwo1           8647.801  21953.435  25329.002  29390.062 39318.85
##    GarageGarSingl     2231.026   4538.113   5867.882   8153.808 11417.84
##    GarageGarDoubl     5302.221  12128.844  15679.468  18372.378 28602.52
##    Tenfree1          -7362.917  -3018.066  -1110.351    655.130  4225.97
##    CenHeat1           6789.054  10087.689  12008.319  13652.972 17671.62
##    AgeBldIntWr          66.841   1648.497   3842.940   8534.553 13274.22
##    AgeBldPostW       -4929.118  -2849.870   -355.913   5982.302 14275.50
##    AgeBld60s        -13142.216  -8457.267  -5520.874  -2885.553  1271.68
##    AgeBld70s        -11356.027  -9566.743  -7285.795  -4238.382  3812.07
##    AgeBld80s         -2849.667   -290.098   2618.808   5548.411 13068.94
##    ************************Diagnostic information*************************
##    Number of data points: 7521
##    Effective number of parameters (2trace(S) - trace(S'S)): 182.7964
##    Effective degrees of freedom (n-2trace(S) + trace(S'S)): 7338.204
##    AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): 175082.7
##    AIC (GWR book, Fotheringham, et al. 2002,GWR p. 96, eq. 4.22): 174949
##    Residual sum of squares: 5479639286184
##    R-square value:  0.5834657
##    Adjusted R-square value:  0.5730883
##
##    ***********************************************************************
##    Program stops at: 2020-05-14 10:15:12
```

The output from the GWR model reveals how the coefficients vary across the 33 boroughs in London. You will see how the global coefficients are exactly the same as the coefficients in the earlier linear model. In this particular model, if we take the area of the floor, we can see that the coefficients range from a minimum value of 612.043 GBP(1 square metre change in the area of the floor resulting in an increase in the average price of the property of 612.043 GBP) to 821.42 GBP(1 square metre change in the area of the floor resulting in an increase in the average price of property of 821.42 GBP). For half of the boroughs in the dataset, as the floor area rises by 1 point, the price of the property will increase between 676.780 GBP and 732.984 GBP(the interquartile range between the 1st Qu and the 3rd Qu).

5.3.2 The Interpretation of Coefficients

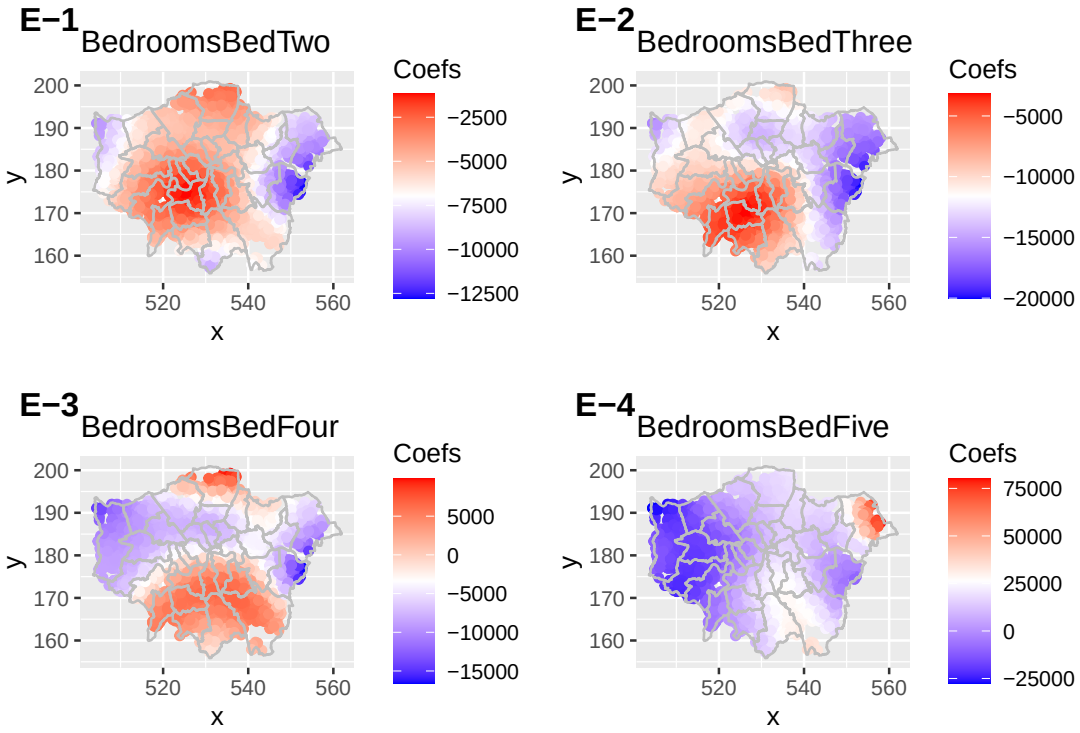
The coefficient ranges can also be seen for the other variables and they suggest some interesting spatial patterning. To explore this we can plot the GWR coefficients for different variables. Firstly we can attach the coefficients to our original dataframe - this can be achieved simply as the coefficients for each ward appear in the same order in our spatial points dataframe as they do in the original dataframe.

- The Interpretation of Coefficient for FlorArea, CenHeat, Bathrooms and Tenfree:



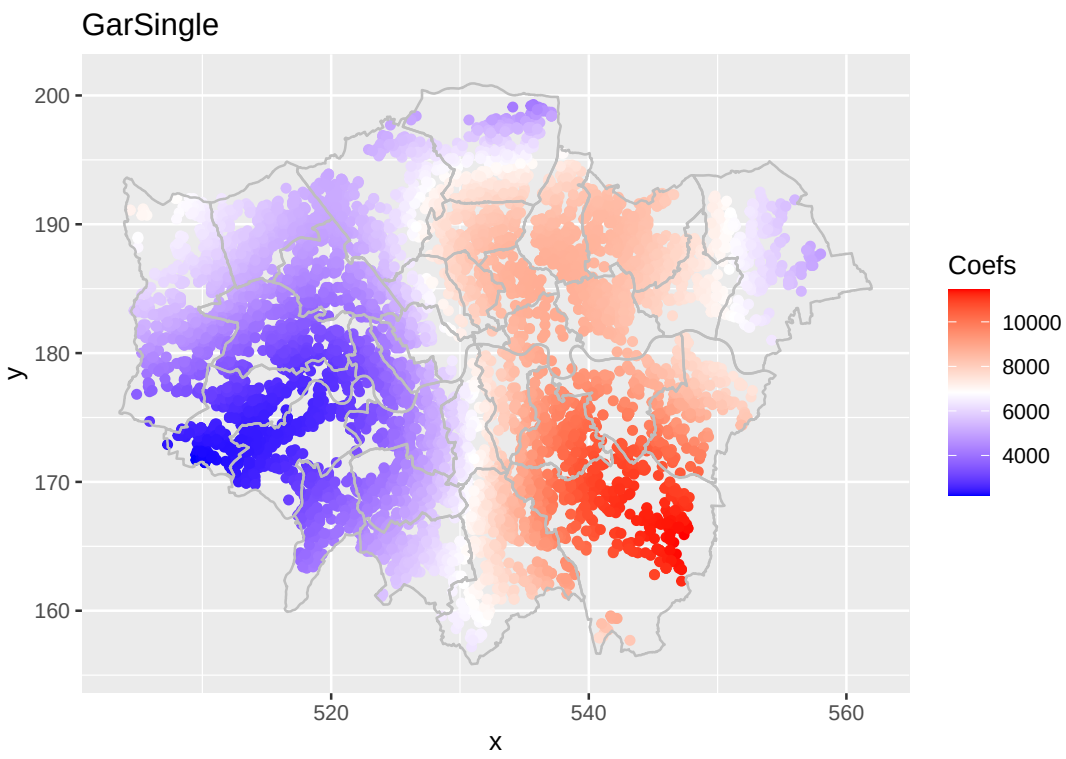
- A: Taking the first plot, which is for the area of floor coefficients. We can see that in the boroughs north of the city center, there is the highest change of property price corresponding to 1 square metre increase. However, in the boroughs south of the city center, the lowest change of property price corresponding to 1 square metre increase. This is a very interesting pattern, but may partly be explained the in the boroughs north of the city center, the buyers value the area of floor much, which makes the area of floor influencing the price of property much.
- B: The second plot is for central heating. In the west and east part of London, having central heating can only influence by less than 10,000 GBP. For those boroughs in the north and south of the city center, the property with central heating is much more important, the price can increase by 12,500 to 17,500 compared with those without central heating.
- C: In this plot, we can see that the price of the property with two or more bathrooms is higher by at least 10,000 GBP than that with one bathroom. Furthermore, in the southwest part of London, the price gap is more significant.
- D: Around City center, the price of freehold property is lower than that of leasehold property. However, in those places far away from the city center, the price of freehold property is higher than that of leasehold property.

- The Interpretation of Coefficient for Bedrooms:

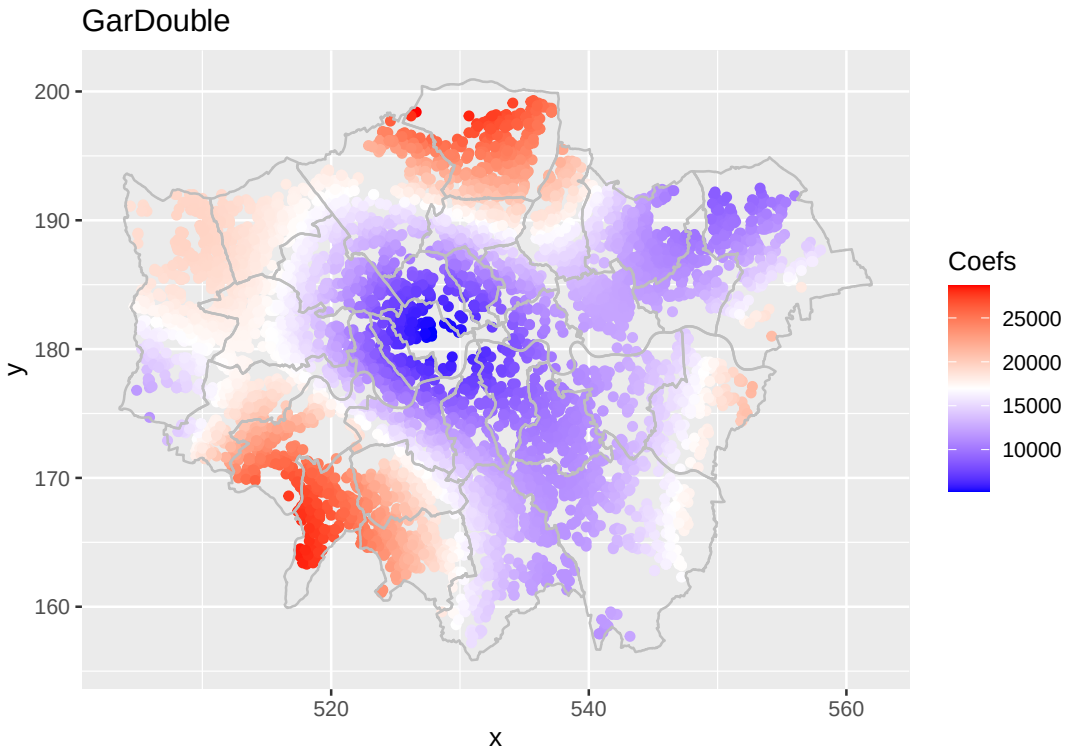


Form the plots E-1, E-2 and E-3, we can see that basically, keeping house area and all other variables constant, the property with more than one bedroom tends to have a lower price than that with one bedroom. The price gap between a property with one bedroom and more bedrooms is larger in boroughs in the southwest part of London than in other places. It should be mention that the number of properties with five bedrooms is the lowest, so the plot E-4 is slightly different from others.

- The Interpretation of Coefficient for Garage:

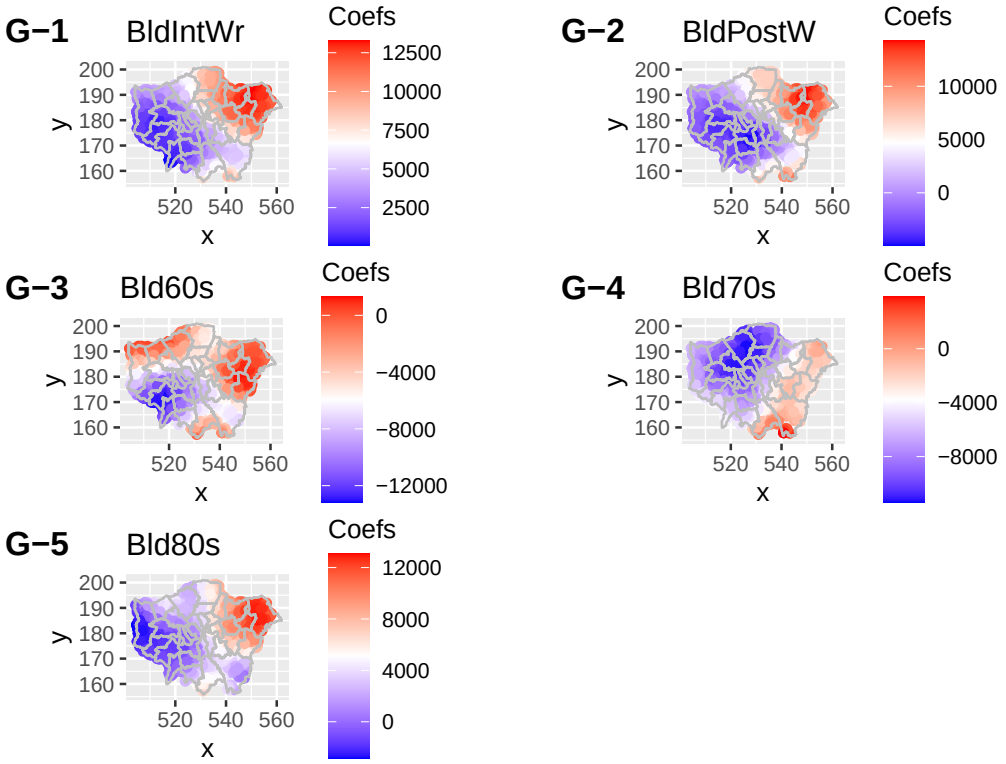


- The Plot shows the prices of properties with one garage is higher than that without garage by about 4,000 GBP to 6,000 GBP in the west part of London. However, in the east part of London, the difference in price is larger, from around 8,000 GBP to 10,000 GBP.



- Compared properties without garage, the properties with two garages are approximate 10,000 GBP to 25,000 GBP. In the very south and north part, the prices are more than 25,000 GBP. Overall, a property with more garages is more expensive.

- The Interpretation of Coefficient for Age:



Overall, compared with properties built before 1914, those built in-war, post-war, and 1980s tend to have a higher price. Especially in the east part of London, the price gap is much larger. However, the price of properties built in the 1960s and 1970s are lower than that built before 1914. In the east part of London, the price gap is larger.

6 Conclusion

In this project, firstly, we clean the data and calculate the predictors based on the original variables. Secondly, analyze the relationship between every variable and target variable property price by point plots to find those variables which have a strong linear relationship with property price. Then, the linear model is built with significant variables, and the variable with high colinearity is moved. Finally, GWR is used to show the influence of the spatial components.

In conclusion, the most reliable determinants on property prices are the area of a floor, the number of bedrooms, having more than two bathrooms, the number of garages, with central heating, Leasehold/Freehold indicator, and the age of properties. Although the global model with these predictors can get a good result for predicting the price of properties, it does not consider the spatial component. It is proved that GWR is a better way to estimate the price of a property.