Executive Summuary

Exploring the anonymised mortgage records data for the London area boroughs for collinearity, adn examining the summaries, scatterplots, and boxplots, appropriate steps can be taking to create a parsimonious model that describes the variability of the data with house price as the response variable and a series of other houseing characteristics as the predictor variables. An Ordinary Least Square (OLS) linear regression model is fitted to the data with optimum predictor variables determined by a stepwise AIC preocedure of the form, Purprice ~ Tenfree + CenHeat + BathTwo + FlorArea + ProfPct + RetiPct + Unemploy + Age + Type + Garage + Bedrooms, which produced an r^2 value of 0.56 and an AIC value of 87571. The OLS model Residuals plot, QQ plot, Scale plot, and Leverage plot are also examined to check how well the model fits the data and determining the model has no non-linear relationships, no normal distribution, homoscedasticity, and no influential outliers. Further exploration is done by examining the median house price, floor area, and residuals for each borough of the London area, this demontrated that some components of the data are best described by local variables via a Geographically Weighted Regression (GWR) model rather than as global variables as in the OLS model. A GWR model is thus created of the same form of the OLS model response and predictor variable relationship, this produced a higher r^2 value of 0.74 and a lower AIC of 86419 value. Therefore the GWR model is better than the OLS model at describing the variablility of the data as it takes into account the spatial heterogeneity, thus is better for house price predictions.

Introduction

In this study anonymised mortgage records data for the London area is examined to determine the best pridictors for a model whose response variable is the purchased price of a house for the boroughs of London. This is done by creating a parsimonious model, a model that describes the variability of the data with as few predictor variables as possible. To that end an Ordinary Least Square (OLS) linear regression model is fitted to the data and examined. Th first step is to check for collinearity in the data, that is when independent variables are highly correlated, this collinearity causes problems by inflation of the variance and loss of procision. By plotting and checking the correlation matrix of the data using corrplot() fuction, variables with potential collinearity problems can be identified and removed. The OLS regression works by minimising the square of the residuals. The fewest predictor variable required to approriately model the data are determined by examining the Akaike Information Criterion (AIC), which examines model quality compared to another model.

The distribution of the pridictor variables over the London boroughs is aslo explored to evaluate how the data changes over location. The data's potential spatial heterogeneity means that using models with global form predictors may not describe the data in its fullest as their relative geographic locations to each other is not taken into account. As such Geographically Weighted Regression (GWR) is also used, this model works by using a moving window weighting technique where the window size is controlled by the bandwidth. A kernal is used to determine the weigthings at each location, starting at the window centre and decayiing as the distance out increases until a set distance is reached. With these geographical weights taken into account, the data can be expressed more locally than globally, and potential provide a better discription of the housing prices in London.

The Data

The data is anonymised mortgage records for the London area, made up of the house price and then a series of property characteristics. Most of the data is comprised of dummy varibles which will be combined into sigle factor variable later.

- Easting - Easting in m
- Northing - Northing in m
- Purprice - Purchase Price in GBP
- BldIntWr - Built between 1918 and 1939
- BldPostW - Built between 1945 and 1959
- Bld60s - Built between 1960 and 1969
- Bld70s - Built between 1970 and 1979
- Bld80s - Built between 1980 and 1989
- TypDetch - Detached property
- TypSemiD - Semi-detached property
- TypFlat - Flat or apartment
- GarSingl - Single Garage
- GarDoubl - Double Garage
- Tenfree - Leasehold/Freehold indicator
- CenHeat - Central heating
- BathTwo - Two or more bathrooms
- BedTwo - Two bedrooms
- BedThree - Three bedrooms
- BedFour - Four bedrooms
- BedFive - Fie bedrooms
- NewPropD - New property
- FlorArea - Floor area in square metres
- NoCarHh - Proportion of households without a car
- CarspP - Cars per person in neighborhood
- ProfPct - Proportion of Households with Professional Head
- UnskPct - Proportion of Households with Unskilled head
- RetiPct - Proportion of residents retired
- Saleunem - Not known
- Unemploy - Unemployed workers
- PopnDnsy - Local population density

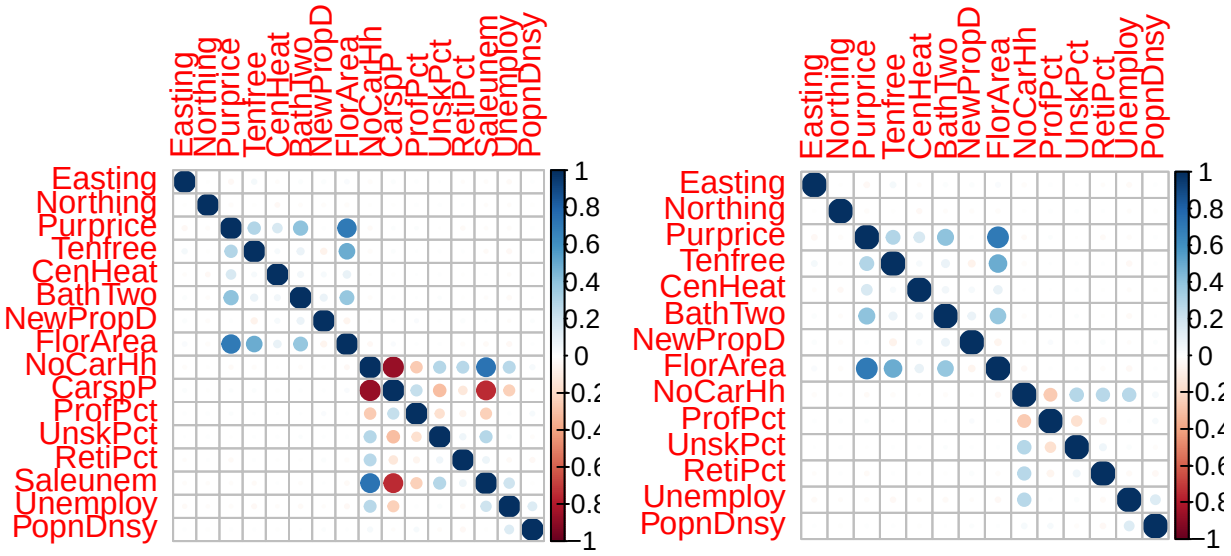
Results

```
# The raw data
LondonData <- read.csv("DataScienceProj.csv",stringsAsFactors=FALSE)

# Function to convert dummy variables to factor variable
Dummy2Factor <- function(mat,lev1="Level1") {
      mat <- as.matrix(mat)
      factor((mat %*% (1:ncol(mat))) + 1,
          labels = c(lev1, colnames(mat)))
}

# Converting dummy variable to factors
Age <- Dummy2Factor(LondonData[,5:9],"PreWW1")
Type <- Dummy2Factor(LondonData[,10:12],"Others")
Garage <- Dummy2Factor(LondonData[,13:14],"HardStnd")
Bedrooms <- Dummy2Factor(LondonData[,18:21],"BedOne")
```

The first thing done was to read in the data and change dummy variables to factors. BldIntWr, BldPostW, Bld60s, Bld70s, and Bld80s where changed into the Age variable with PreWW1 accounting for the unlabelled. TypDetch, TypSemiD, and TypFlat where changed into Type. GarSingl, GarDoubl, and HardStnd make up

the variable Garage. Finally, BedOne, BedTwo, BedThree, BedFour, and BedFive make up the Bedrooms variable.

```
par(mfrow=c(1,2))
LondonDataNew <- data.frame(LondonData[,c(2:4,15:17,22,23:31)],Age,Type,Garage,Bedrooms) # New data
corrplot::corrplot(cor(LondonDataNew[,-c(17:20)])) # Plot of correlation matrix of new data

LondonDataNew <- data.frame(LondonData[,c(2:4,15:17,22,23:24,26:28,30:31)],Age,Type,Garage,Bedrooms) # 
corrplot::corrplot(cor(LondonDataNew[,-c(15:18)])) # Plot of new correlation matrix
```




```
par(mfrow=c(1,1))
```

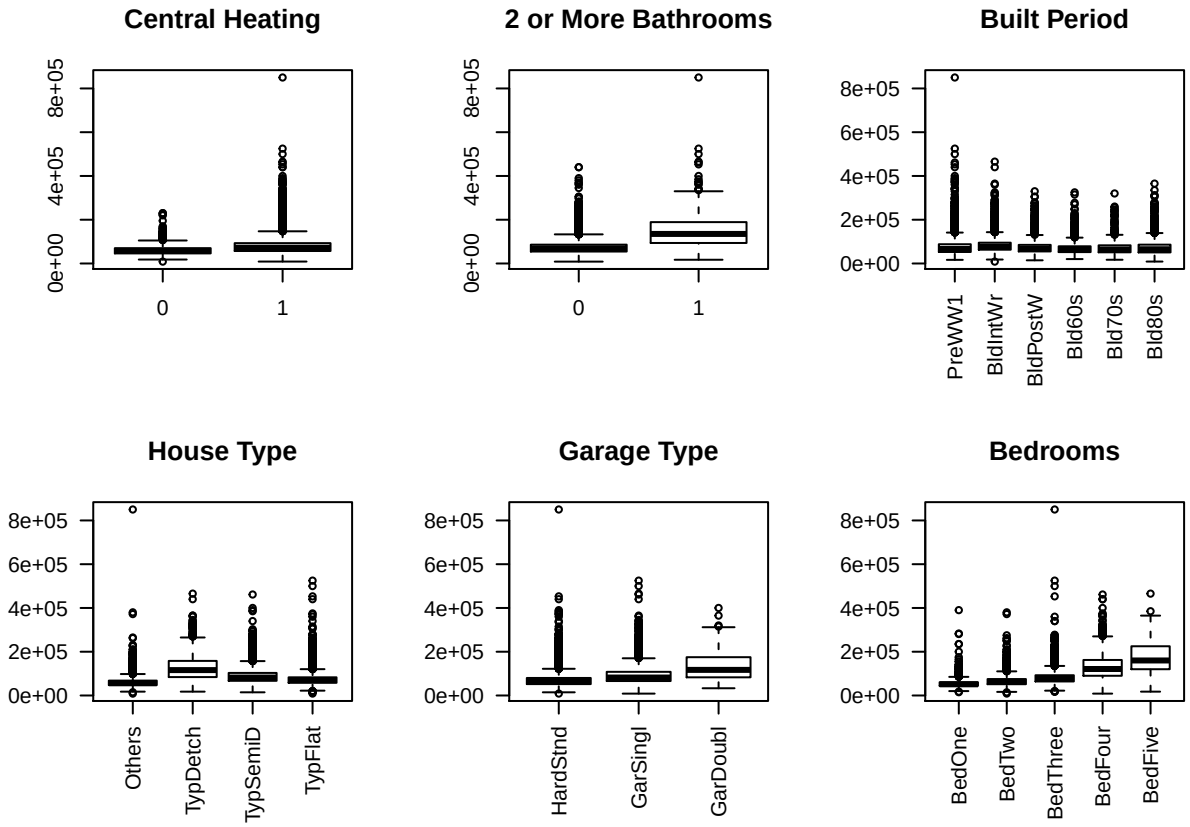
By plotting the correlation matrix variables with collinearity issues become identifiable. The variable CarspP has a strong correlation with both the NoCarHh variable and the Saleunem variable. As such CarspP was removed to reduce the collinearty problem. The variable Saleunem was also removed as the describtion of the variable is not known.

```
summary(LondonDataNew)
```

```
##     Easting          Northing        Purprice         Tenfree      
##  Min.   :504400   Min.   :157200   Min.   :  8500   Min.   :0.0000  
##  1st Qu.:517800   1st Qu.:172700   1st Qu.: 55000   1st Qu.:0.0000  
##  Median :527600   Median :181200   Median : 70000   Median :1.0000  
##  Mean   :527926   Mean   :180009   Mean   : 80018   Mean   :0.6835  
##  3rd Qu.:536700   3rd Qu.:187400   3rd Qu.: 90000   3rd Qu.:1.0000  
##  Max.   :558000   Max.   :200100   Max.   :850000   Max.   :1.0000  
##     CenHeat          BathTwo           NewPropD          FlorArea      
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   : 23.22  
##  1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.: 71.77  
##  Median :1.0000   Median :0.00000   Median :0.00000   Median : 91.02  
##  Mean   :0.8789   Mean   :0.05392   Mean   :0.03638   Mean   : 96.49  
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:112.11  
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :278.00  
##     NoCarHh          ProfPct          UnskPct          RetiPct      
##  Min.   : 0.00   Min.   :  0.000   Min.   : 0.000   Min.   :  0.00  
##  1st Qu.:14.81   1st Qu.:  0.000   1st Qu.: 0.000   1st Qu.: 18.75  
##  Median :26.96   Median :  5.556   Median : 0.000   Median : 35.29  
##  Mean   :29.26   Mean   :  7.640   Mean   : 4.216   Mean   : 45.95
```

```
##  3rd Qu.:41.26   3rd Qu.: 12.500   3rd Qu.: 7.692   3rd Qu.: 58.33
##  Max.   :92.62   Max.   :100.000   Max.   :71.429   Max.   :900.00
##     Unemploy         PopnDnsy           Age             Type
##  Min.   :  0.00   Min.   : 0.000   PreWW1  :4261   Others  :3791
##  1st Qu.: 15.07   1st Qu.: 5.296   BldIntWr:4365   TypDetch:1168
##  Median : 38.86   Median : 7.623   BldPostW:1054   TypSemiD:3260
##  Mean   : 47.18   Mean   : 8.882   Bld60s  : 789   TypFlat :4317
##  3rd Qu.: 64.90   3rd Qu.:10.870   Bld70s  : 679
##  Max.   :686.99   Max.   :82.803   Bld80s  :1388
##       Garage         Bedrooms
##  HardStnd:8306   BedOne  :1713
##  GarSingl:3923   BedTwo  :3785
##  GarDoubl: 307   BedThree:5723
##                  BedFour :1100
##                  BedFive : 215
##
```

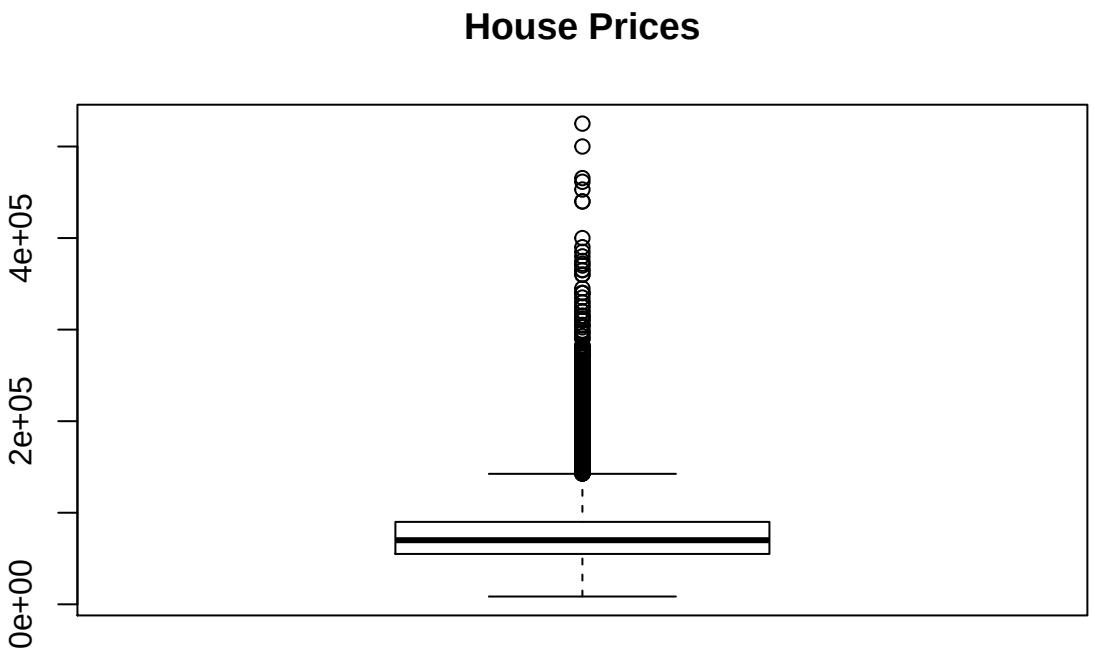
```
# Boxplots of data
par(mfrow=c(2,3))
boxplot(Purprice~CenHeat,data=LondonDataNew, main="Central Heating")
boxplot(Purprice~BathTwo,data=LondonDataNew, main="2 or More Bathrooms")
boxplot(Purprice~Age,data=LondonDataNew, main="Built Period",las=2)
boxplot(Purprice~Type,data=LondonDataNew, main="House Type",las=2)
boxplot(Purprice~Garage,data=LondonDataNew, main="Garage Type",las=2)
boxplot(Purprice~Bedrooms,data=LondonDataNew, main="Bedrooms",las=2)
```



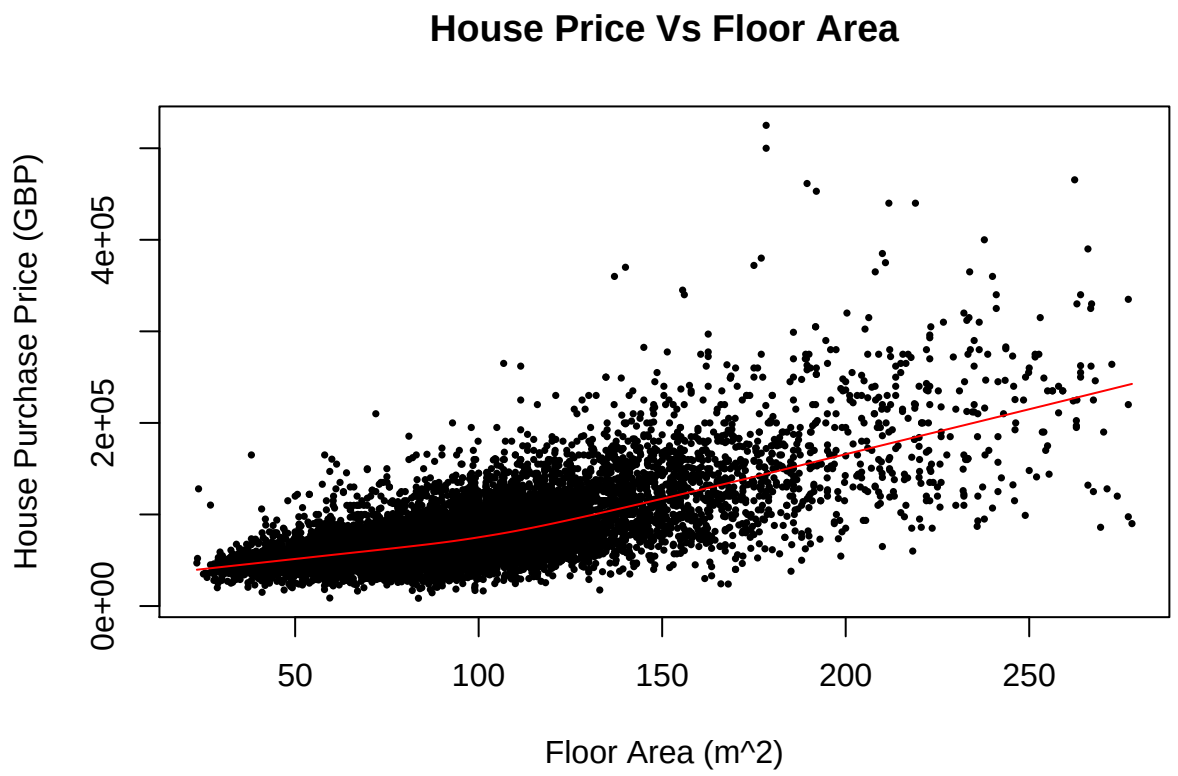

```
par(mfrow=c(1,1))
```

By examining the summary and boxplots data that the medians of the factor variables are quite similar to each other, and that the ranges for the data are with reasonable expectations except for RetoPct and Unemploy whose Max values seem to have taken a drastic jumo. This could be due to some spatial heterogeneity, with some boroughs having relatively high Retirement or Unemployment numbers.

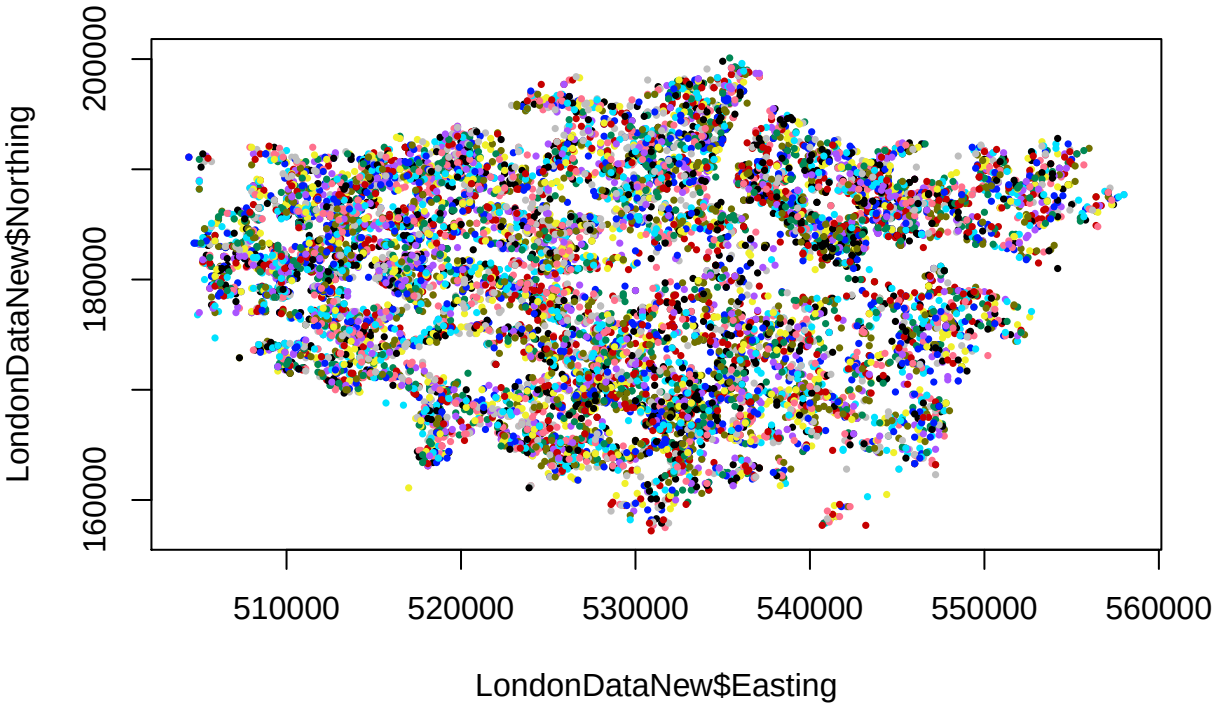
```
# Boxplot of house price data
LondonDataNew <- LondonDataNew[LondonDataNew$Purprice < 600000,]
boxplot(LondonDataNew$Purprice, main="House Prices")
```




```
# Hosues price vs Floor Area plot
plot(LondonDataNew[,c("FlorArea","Purprice")],pch=16,cex=0.5, main="House Price Vs Floor Area", xlab="Fl
lines(lowess(LondonDataNew[,c("FlorArea","Purprice")]),col="red")
```




```
# Map of house prices on by eastings and northings
Classes <- classIntervals(LondonDataNew$Purprice,10,"quantile")
Colours <- findColours(Classes,palette())
plot(LondonDataNew$Easting,LondonDataNew$Northing,pch=16,cex=0.5,col=Colours)
```




```
# Coordinate trends
CoordMod <- lm(Purprice~Easting+Northing+(Easting^2)+(Northing^2)+(Easting*Northing),
```

```
            data=LondonDataNew)
summary(CoordMod)
```

```
##
## Call:
## lm(formula = Purprice ~ Easting + Northing + (Easting^2) + (Northing^2) +
##     (Easting * Northing), data = LondonDataNew)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -72630 -24983 -10067   9733 443497
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -5.172e+05  4.163e+05  -1.242    0.214
## Easting           1.152e+00  7.850e-01   1.467    0.142
## Northing          3.679e+00  2.290e+00   1.607    0.108
## Easting:Northing -7.083e-06  4.317e-06  -1.641    0.101
##
## Residual standard error: 41090 on 12531 degrees of freedom
## Multiple R-squared:  0.002068,   Adjusted R-squared:  0.001829
## F-statistic: 8.656 on 3 and 12531 DF,  p-value: 9.813e-06
```

By examining the house price verses the floor area of the house it can be seen that as the house size increases so does the price. However there is still many outliers as evident by the boxplot, why? Other variables might be required to account for the discrepency as house price isn't soley depenedent on its size or there are potentially geographic reasons as houses in one borough may not cost the same as houses in another borough even if they are the same size. The map of the house prices over the eastings and northings shows just how distributed the prices are. By applying a model relating the house price to the eastings and northings, in this case a quadratic model had the lowest AIC, it can be seen that the prices get higher by moving north and west, while they get lower by moving south and east. Further prdictor variable are required to account for discrpency of the above relationship.

```
# The Model
price.lm   <- lm(Purprice~., data=LondonDataNew[,-c(1,2)]) # Making model without easting and northing
invisible(capture.output(price.step <- stepAIC(price.lm))) # Determining which variables to keep
summary(price.step)
```

```
##
## Call:
## lm(formula = Purprice ~ Tenfree + CenHeat + BathTwo + FlorArea +
##     ProfPct + RetiPct + Unemploy + Age + Type + Garage + Bedrooms,
##     data = LondonDataNew[, -c(1, 2)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -137150  -13512   -1351   10307  371750
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5950.965   1253.518   4.747 2.08e-06 ***
## Tenfree            6187.420   1351.774   4.577 4.76e-06 ***
## CenHeat           11863.252    754.439  15.725  < 2e-16 ***
## BathTwo           24077.868   1202.238  20.028  < 2e-16 ***
## FlorArea            677.868     11.377  59.584  < 2e-16 ***
```

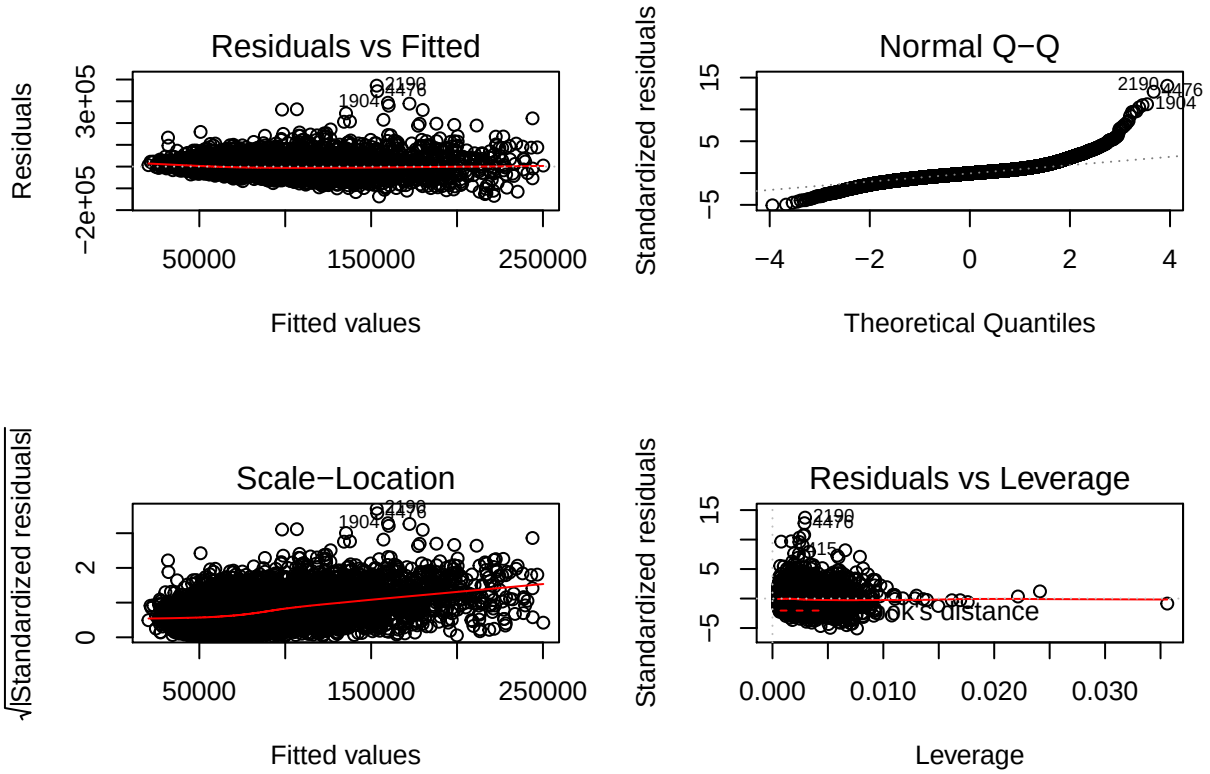
```
## ProfPct              42.865     23.894   1.794 0.072842 .
## RetiPct              -7.637      5.198  -1.469 0.141785
## Unemploy             11.907      5.476   2.174 0.029697 *
## AgeBldIntWr        4019.195    656.810   6.119 9.68e-10 ***
## AgeBldPostW       -1113.333    974.911  -1.142 0.253482
## AgeBld60s         -7299.877   1089.574  -6.700 2.18e-11 ***
## AgeBld70s         -6701.010   1164.197  -5.756 8.82e-09 ***
## AgeBld80s           970.234    898.802   1.079 0.280397
## TypeTypDetch       5796.442   1657.973   3.496 0.000474 ***
## TypeTypSemiD      -6641.390   1440.581  -4.610 4.06e-06 ***
## TypeTypFlat      -11557.060   1395.549  -8.281  < 2e-16 ***
## GarageGarSingl     3750.774    614.425   6.105 1.06e-09 ***
## GarageGarDoubl     9236.796   1675.834   5.512 3.62e-08 ***
## BedroomsBedTwo    -3419.603    868.894  -3.936 8.34e-05 ***
## BedroomsBedThree  -7882.546   1067.921  -7.381 1.67e-13 ***
## BedroomsBedFour   -1731.162   1541.680  -1.123 0.261499
## BedroomsBedFive    3941.049   2503.710   1.574 0.115493
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27150 on 12513 degrees of freedom
## Multiple R-squared:  0.565,  Adjusted R-squared:  0.5643
## F-statistic:   774 on 21 and 12513 DF,  p-value: < 2.2e-16
```

```
price.step$anova # Model comparison
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Purprice ~ Tenfree + CenHeat + BathTwo + NewPropD + FlorArea +
##     NoCarHh + ProfPct + UnskPct + RetiPct + Unemploy + PopnDnsy +
##     Age + Type + Garage + Bedrooms
##
## Final Model:
## Purprice ~ Tenfree + CenHeat + BathTwo + FlorArea + ProfPct +
##     RetiPct + Unemploy + Age + Type + Garage + Bedrooms
##
##
##         Step Df   Deviance Resid. Df   Resid. Dev      AIC
## 1                              12509 9.218278e+12 255965.7
## 2  - NoCarHh  1  538603718     12510 9.218817e+12 255964.4
## 3  - UnskPct  1  438795881     12511 9.219256e+12 255963.0
## 4 - PopnDnsy  1  749284811     12512 9.220005e+12 255962.0
## 5 - NewPropD  1 1128137530     12513 9.221133e+12 255961.6
```

A linear model is now created of the form, Purprice ~ Tenfree + CenHeat + BathTwo + NewPropD + FlorArea + NoCarHh + ProfPct + UnskPct + RetiPct + Unemploy + PopnDnsy + Age + Type + Garage + Bedrooms, using the stepAIC() function a stepwise procedure removes predictor variables on their AIC. From the model summary and the anova table it can be seen that the final model is Purprice ~ Tenfree + CenHeat + BathTwo + FlorArea + ProfPct + RetiPct + Unemploy + Age + Type + Garage + Bedrooms, and it has an r^2 value of 0.565.

```
par(mfrow=c(2,2))
plot(price.step)
```

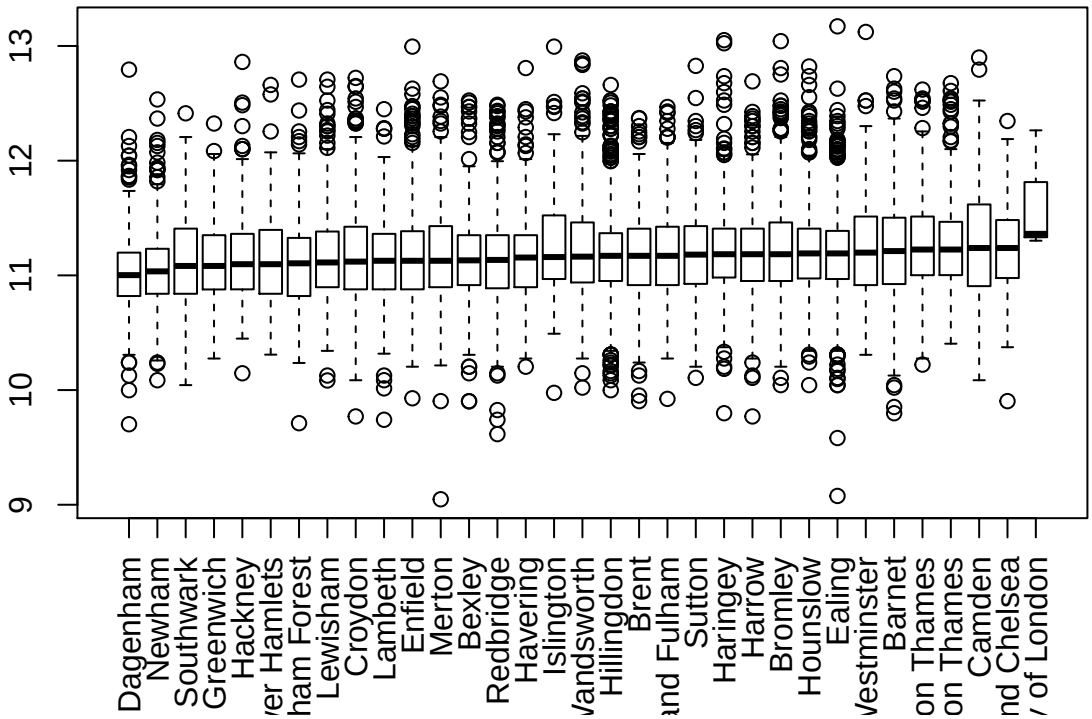



```
par(mfrow=c(1,1))
```

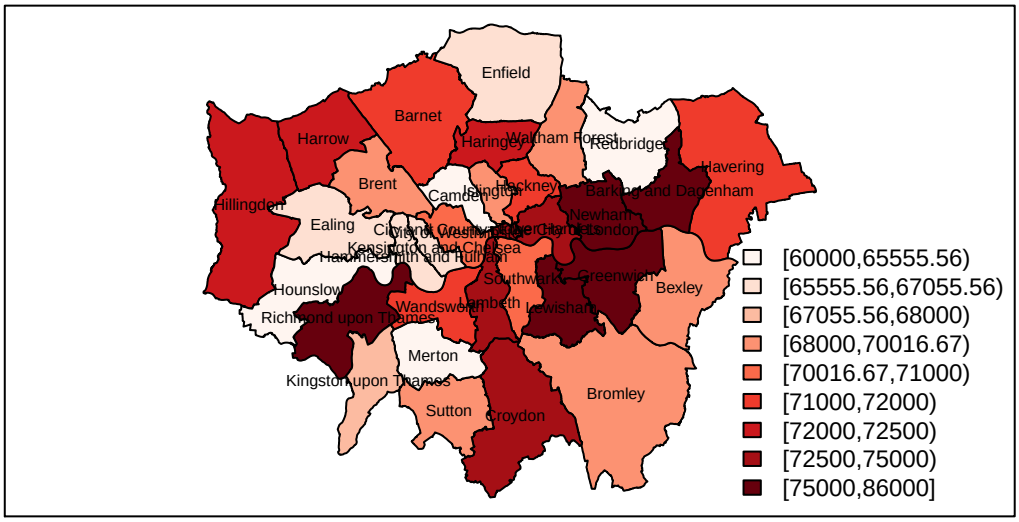
To further examine if this model is a good fit for the data Residual, QQ, Scale, and Leverage plots can be explored. The Residuals verses Fitted values plot shows if the residuals had any non-linear patterns that the model didn't capture, however since the residuals are equally spread around the horizontal line without distinct patterns there is probably no non-linear relationships. The Normal Q-Q plot shows if the residuals are normally distributed, and while there are along the normal line in the middle they also have heavy tails, thus the data probably isn't normally distributed.

To examine homoscedasticity the Scale-Location plot can be used as it shows if the residuals are spread equally along the predictors. While not completely horizontal, the residuals do have some equal variance. Finally, the Residuals verses Leverage determines wether there is influential outliers, in this case there is not as there is no visiable Cook's distance lines.

```
# Variation by borough
invisible(capture.output(LB <- readOGR(dsn=".",layer="LondonBoroughs",stringsAsFactors=FALSE)))  # Load
LH <- SpatialPointsDataFrame(LondonDataNew[,1:2],LondonDataNew) # Makeing SPDF
proj4string(LH) <- CRS(proj4string(LB)) # copy CRS
LHLB <- over(LH,LB)   # spatial joining points and polygons
LondonDataNew$Borough <- gsub(" London Boro","",LHLB$NAME)  # add borough names only to data
Boroughs <- names(table(LondonDataNew$Borough)) # borough names

# Price by borough
b.order <- rank(tapply(LondonDataNew$Purprice+runif(nrow(LondonDataNew)),
                       LondonDataNew$Borough,median)) # ranking boroughs by the median of prices

boxplot(log(Purprice)~Borough,data=LondonDataNew,xaxt="n", at=b.order)
axis(1,labels=Boroughs,at=b.order,las=2)
title("Log of Price by Borough")
```



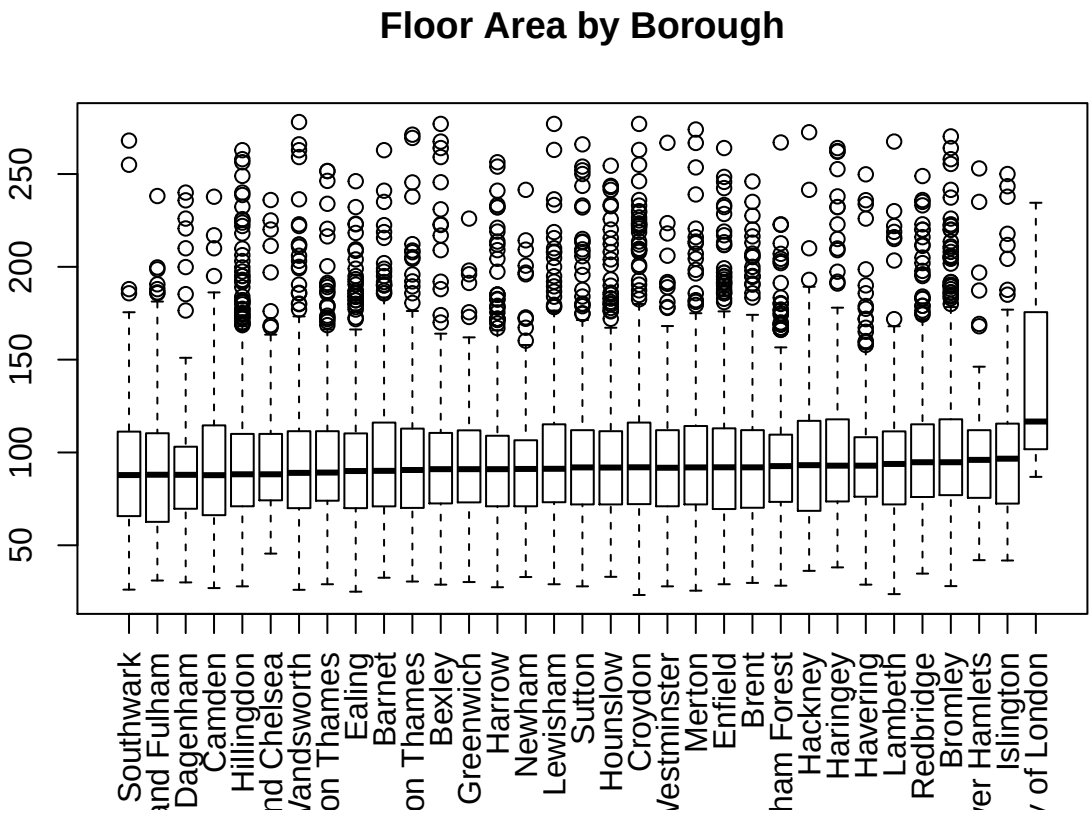

```
quickMap2(tapply(LondonDataNew$Purprice,LondonDataNew$Borough,median),plotNames=TRUE,dp=3)
```



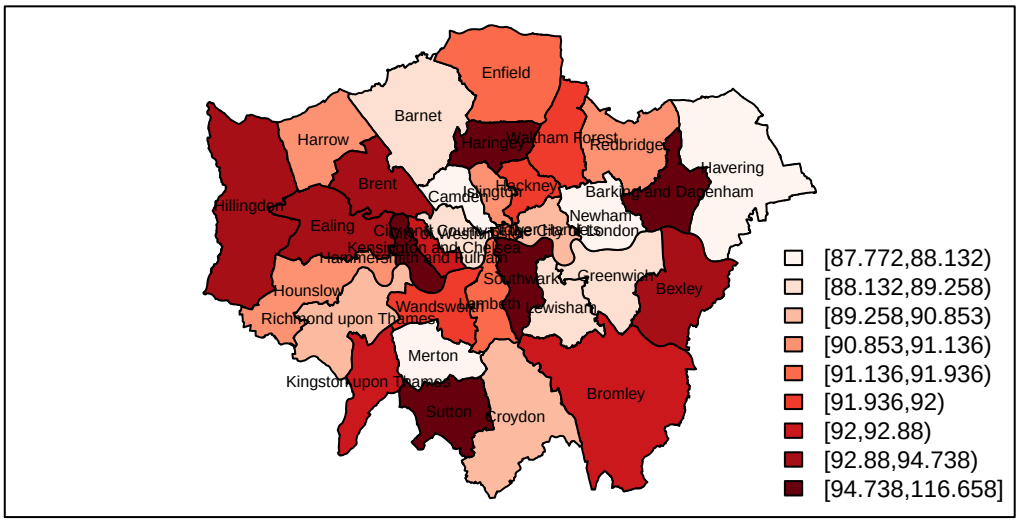
To examine if location may play a role the median housing prices where expressed by their boroughs. As seen in the boxplots their median are quite similar to each other, with only the City of London showing a trend of higher house prices. The borough plot shows that there is a fairly even spread of house prices accross London with the prices seeming to increase the closer to Londons centre the borough is. While the median house price deviations over the geographical location may be small, it's still present and can potential describe the variablity of the data a little bit better.

```
# Floor area by borough
b.order.floor <- rank(tapply(LondonDataNew$FlorArea+runif(nrow(LondonDataNew)),
                      LondonDataNew$Borough,median)) # ranking boroughs by the median of floor area
```

```
boxplot(FlorArea~Borough,data=LondonDataNew,xaxt="n", at=b.order.floor)
axis(1,labels=Boroughs,at=b.order.floor,las=2)
title("Floor Area by Borough")
```

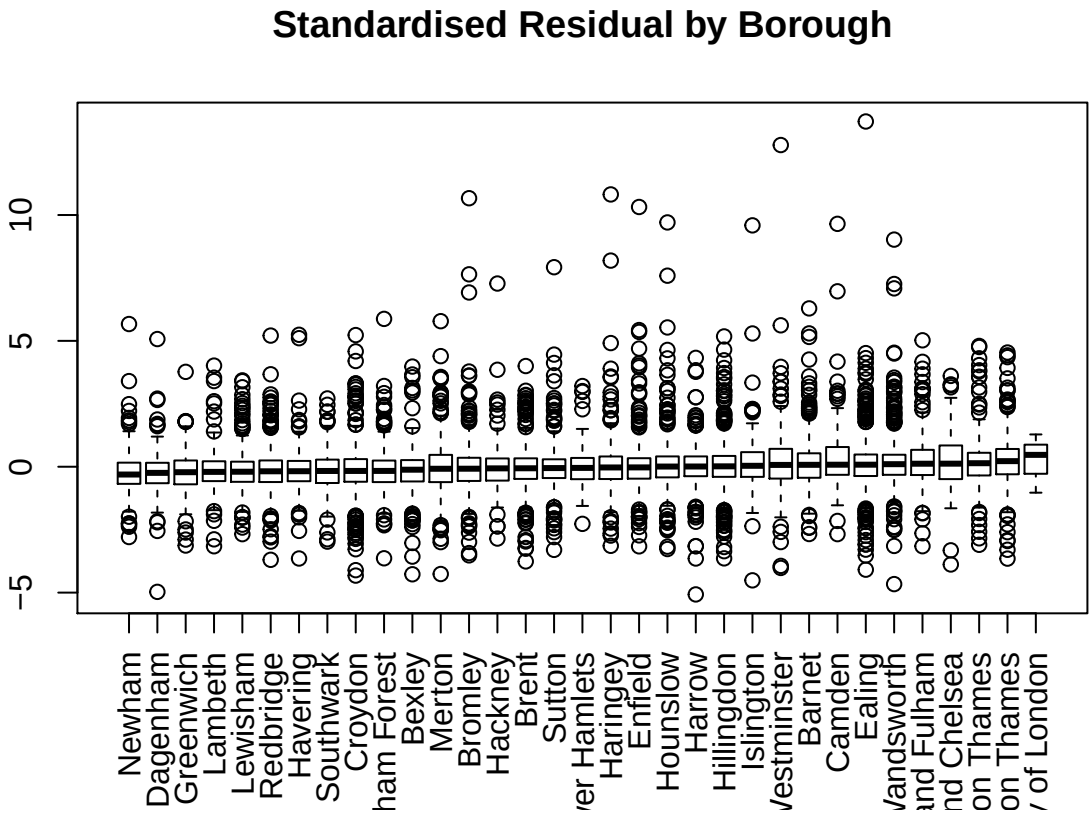



```
quickMap2(tapply(LondonDataNew$FlorArea,LondonDataNew$Borough,median),plotNames=TRUE,dp=3)
```

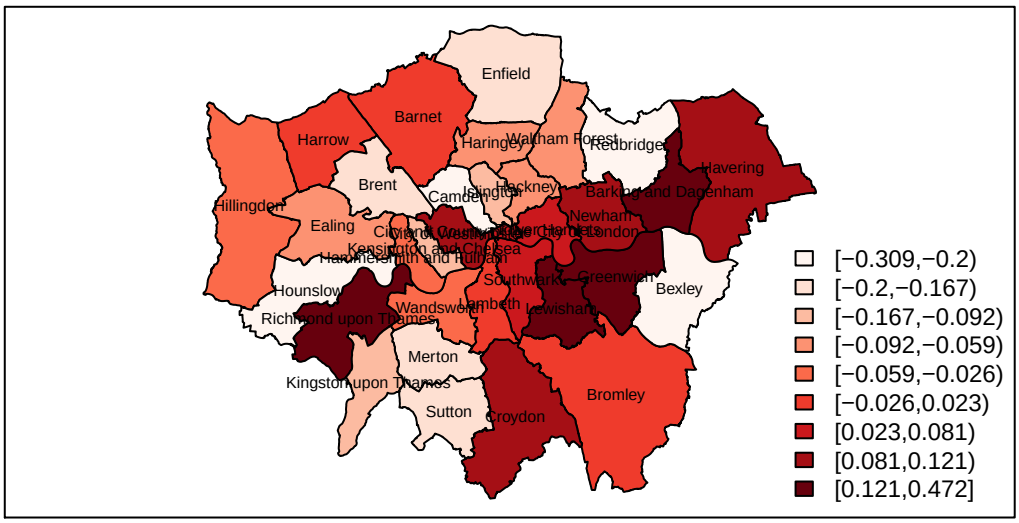


Further examination into median variable distributions over it's respective borough shows that the median floor area also has a slight change from borough to borough. While the boxplot shows again that the median values are not too dissimilar, except for the City of London, the borough plot again shows that the difference is there. The borough plot shows that the median floor area gets lower the closer to London centre the borough is, again demonstrating that a more local model might be a better fit.

```
# Residuals by borough
LondonDataNew$stdres.price.step <- stdres(price.step)
b.order.price.step <- rank(tapply(LondonDataNew$stdres.price.step+
                                  runif(nrow(LondonDataNew))*0.0001,LondonDataNew$Borough,median))

boxplot(stdres.price.step~Borough, data=LondonDataNew, xaxt="n", at=b.order.price.step)
axis(1,labels=Boroughs,at=b.order.price.step,las=2)
title("Standardised Residual by Borough")
```




```
quickMap2(tapply(LondonDataNew$stdres.price.step,LondonDataNew$Borough,median),plotNames=TRUE,dp=3)
```



Finally, the boxplot of the median residuals by borough again shows that the median residual values dont change drastically from one borough to the other. The borough plot shows that the median residuals get larger the more eastern the borough is.

```
# GWR
set.seed(1)
s <- sample(nrow(LondonDataNew), round(.3*nrow(LondonDataNew))) # Splitting data
LondonDataTrain <- LondonDataNew[s,] # Training set
LondonDataTest <- LondonDataNew[-s,] # Testing set

LondonDataTrain <- LondonDataTrain[,-c(19:20)] # Removing un-needed variable
LondonDataTrain <- SpatialPointsDataFrame(cbind(LondonDataTrain[,1:2]),LondonDataTrain) # Making SPDF
gwr<-gwr.basic(Purprice ~ Tenfree + CenHeat + BathTwo + FlorArea + ProfPct + RetiPct +
                 Unemploy + Age + Type + Garage + Bedrooms,
               data=LondonDataTrain, bw=250, adaptive=T)
gwr
```

```
##    ***********************************************************************
##    *                       Package   GWmodel                             *
##    ***********************************************************************
##    Program starts at: 2019-05-16 12:37:56
##    Call:
##    gwr.basic(formula = Purprice ~ Tenfree + CenHeat + BathTwo +
##     FlorArea + ProfPct + RetiPct + Unemploy + Age + Type + Garage +
##     Bedrooms, data = LondonDataTrain, bw = 250, adaptive = T)
##
##    Dependent (y) variable:  Purprice
##    Independent variables:  Tenfree CenHeat BathTwo FlorArea ProfPct RetiPct Unemploy Age Type Garage
##    Number of data points: 3760
##    ***********************************************************************
##    *                    Results of Global Regression                     *
##    ***********************************************************************
##
##    Call:
##     lm(formula = formula, data = data)
##
##    Residuals:
##     Min      1Q  Median      3Q     Max
## -115671  -13618   -1411   10478  345044
##
##    Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
##    (Intercept)       3.037e+03  2.351e+03   1.292 0.196566
##    Tenfree           8.650e+03  2.466e+03   3.508 0.000456 ***
##    CenHeat           1.276e+04  1.409e+03   9.057  < 2e-16 ***
##    BathTwo           2.013e+04  2.253e+03   8.938  < 2e-16 ***
##    FlorArea          7.350e+02  2.075e+01  35.415  < 2e-16 ***
##    ProfPct          -1.329e+01  4.438e+01  -0.299 0.764674
##    RetiPct           3.856e+00  9.624e+00   0.401 0.688719
##    Unemploy         -2.162e-01  9.862e+00  -0.022 0.982516
##    AgeBldIntWr       4.303e+03  1.219e+03   3.529 0.000422 ***
##    AgeBldPostW       1.037e+03  1.762e+03   0.589 0.556170
##    AgeBld60s        -6.073e+03  1.990e+03  -3.052 0.002291 **
##    AgeBld70s        -8.777e+03  2.192e+03  -4.005 6.32e-05 ***
##    AgeBld80s         1.584e+03  1.671e+03   0.948 0.343136
```

```
##    TypeTypDetch      1.403e+03  3.023e+03   0.464 0.642629
##    TypeTypSemiD     -9.306e+03  2.629e+03  -3.539 0.000406 ***
##    TypeTypFlat      -1.379e+04  2.549e+03  -5.412 6.61e-08 ***
##    GarageGarSingl    3.962e+03  1.138e+03   3.480 0.000506 ***
##    GarageGarDoubl    3.294e+03  3.078e+03   1.070 0.284687
##    BedroomsBedTwo   -5.456e+03  1.601e+03  -3.408 0.000661 ***
##    BedroomsBedThree -1.085e+04  1.938e+03  -5.597 2.34e-08 ***
##    BedroomsBedFour  -6.186e+03  2.829e+03  -2.187 0.028809 *
##    BedroomsBedFive  -5.876e+03  4.899e+03  -1.200 0.230393
##
##    ---Significance stars
##    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##    Residual standard error: 27530 on 3738 degrees of freedom
##    Multiple R-squared: 0.5694
##    Adjusted R-squared: 0.567
##    F-statistic: 235.4 on 21 and 3738 DF,  p-value: < 2.2e-16
##    ***Extra Diagnostic information
##    Residual sum of squares: 2.832697e+12
##    Sigma(hat): 27455.01
##    AIC:  87571.1
##    AICc:  87571.4
##    ***************************************************************************
##    *          Results of Geographically Weighted Regression                  *
##    ***************************************************************************
##
##    ********************Model calibration information*********************
##    Kernel function: bisquare
##    Adaptive bandwidth: 250 (number of nearest neighbours)
##    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.
##    Intercept        -6.2594e+04 -6.6670e+03  4.4568e+03  1.5676e+04
##    Tenfree          -4.1631e+04 -2.2837e+03  4.3880e+03  1.1060e+04
##    CenHeat          -4.0712e+03  6.5475e+03  1.1927e+04  1.7974e+04
##    BathTwo          -4.6353e+04  3.9343e+03  1.9035e+04  3.4684e+04
##    FlorArea          2.2750e+02  5.4154e+02  6.8798e+02  8.2988e+02
##    ProfPct          -6.1894e+02 -8.6886e+01  1.3320e+01  1.3860e+02
##    RetiPct          -1.5954e+02 -2.0774e+01  6.1752e+00  3.4066e+01
##    Unemploy         -1.3400e+02 -3.2836e+01 -1.9861e+00  2.7410e+01
##    AgeBldIntWr      -1.5827e+04 -1.9436e+02  4.1939e+03  9.4553e+03
##    AgeBldPostW      -2.4772e+04 -4.8049e+03  9.9935e+02  7.5837e+03
##    AgeBld60s        -3.9698e+04 -1.2290e+04 -6.2559e+03  1.8912e+03
##    AgeBld70s        -7.7777e+04 -1.4909e+04 -7.7092e+03 -1.9195e+03
##    AgeBld80s        -3.4559e+04 -3.2613e+03  2.4192e+03  7.7873e+03
##    TypeTypDetch     -5.1909e+04 -2.6724e+03  7.0168e+03  1.6384e+04
##    TypeTypSemiD     -6.1405e+04 -1.2197e+04 -4.1010e+03  3.2719e+03
##    TypeTypFlat      -7.1401e+04 -1.4865e+04 -8.3841e+03 -2.7045e+03
##    GarageGarSingl   -1.0693e+04 -3.4828e+02  4.1160e+03  8.1743e+03
##    GarageGarDoubl   -7.3219e+04 -1.3874e+04  3.2075e+03  1.9286e+04
##    BedroomsBedTwo   -2.9237e+04 -7.5341e+03 -3.7983e+03  1.3267e+02
##    BedroomsBedThree -3.7208e+04 -1.6029e+04 -7.9777e+03 -1.8890e+03
##    BedroomsBedFour  -6.4028e+04 -1.5778e+04 -5.7475e+03  7.0378e+03
```

```
##    BedroomsBedFive  -2.1255e+05 -3.0179e+04 -1.8595e+03  2.9463e+04
##                          Max.
##    Intercept         47488.46
##    Tenfree           61570.50
##    CenHeat           33481.16
##    BathTwo          123231.90
##    FlorArea           1338.34
##    ProfPct             621.57
##    RetiPct             203.59
##    Unemploy            173.18
##    AgeBldIntWr       20243.63
##    AgeBldPostW       27499.17
##    AgeBld60s         49609.17
##    AgeBld70s         35982.54
##    AgeBld80s         25286.25
##    TypeTypDetch      44298.72
##    TypeTypSemiD      34226.46
##    TypeTypFlat       30988.54
##    GarageGarSingl    26255.54
##    GarageGarDoubl    73134.07
##    BedroomsBedTwo    19540.27
##    BedroomsBedThree  20046.54
##    BedroomsBedFour   63416.12
##    BedroomsBedFive  155674.25
##    ************************Diagnostic information*************************
##    Number of data points: 3760
##    Effective number of parameters (2trace(S) - trace(S'S)): 1064.508
##    Effective degrees of freedom (n-2trace(S) + trace(S'S)): 2695.492
##    AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): 87701.39
##    AIC (GWR book, Fotheringham, et al. 2002,GWR p. 96, eq. 4.22): 86419.18
##    Residual sum of squares: 1.6971e+12
##    R-square value:  0.742037
##    Adjusted R-square value:  0.640124
##
##    ***********************************************************************
##    Program stops at: 2019-05-16 12:39:39
```

Taking into account the previous few graphs that show the distribution of median values for house price, floor area, and residuals over their respective boroughs, and taking into the account that given the data available the most optimum OLS regression model was made with an r^2 value of 0.56, the next solution to describe more of the variability of the data is Geographically Weigthed Regression. Given the size of the data it was first broken into a training set and a testing set, with the training set being made up of 3760 rows. The data was then transformed into a spatial points data frame, using the eastings and northings, and a regression model made with gwr.basic() function from the GWmodels package. This model does in fact describe more of the data, with an r^2 value of 0.74, along with a lower AIC of 86419 compared to that of 87571 for the OLS model.

Conclusion

Utilinsing the stepwise AIC procedure the optimum predictors, given the data, for a parsimonious model to predict housing price where determined to be Tenfree, CenHeat, BathTwo, FlorArea, ProfPct, RetiPct, Unemploy, Age, Type, Garage, and Bedrooms. The OLS linear regression model for those predictor had an r^2 value of 0.56, with no non-linear relationships, no normal distribution, homoscedasticity, and no influential outliers. However, while the OLS model is good, it can be imporved by taking into account the data's spatial heterogeneity as seen in the borough plots above. This improvement is done by using the same predictor, but expressing them as local variabl over a given range rather than globale variables. This new geographically weigthed regression model produces a higher r^2 value of 0.74 and a lower AIC of 86419 value, meaning that the GWR model describe more of the variability of the data and is a better model for prdicting house prices.