House Price Prediction

The data file "House.csv" contains information on the sale of 76 single-family homes in Dublin during 2005. We will model single-family home sale price by ($\frac{\text{Price in thousands of eur}}{100{,}000}$), which range from 155.5 (€155,500) to 450.0 (€450,000), using these predictor variables:

Size floor size (thousands of square feet)

Lot lot size category (from 1 to 11 explained below)

Bath number of bathrooms (with half-bathrooms counting as 0.1 explained below)

Bed number of bedrooms (between 2 and 6)

Year year the house was built.

Garage garage size (0, 1, 2, or 3 cars)

High indicator for The High School (reference: The High School)

Alexandra indicator for Alexandra College (reference: Alexandra College)

Stratford indicator for Stratford College (reference: Stratford College)

St.Mary's indicator for St.Mary's College (reference: St.Mary's College)

St Louis indicator for St Louis High School (reference: St Louis High School)

It seems reasonable to expect that homes built on properties with a large amount of land area command higher sale prices than homes with less land, all else being equal. However, an increase in land area of (say) 2000 square feet from 4000 to 6000 should probably make a larger difference (to sale price) than going from 24,000 to 26,000. Thus, realtors have constructed lot size "categories," which in their experience correspond to approximately equal-sized increases in sale price.

Lot Size	0-3000ft^2	3000-5000ft^2	5000-7000ft^2	7000-10,000ft^2
Category	1	2	3	4
Lot Size	10,000-15,000ft^2	15,000-20,000ft^2	20,000ft^2-1 acre	1-3 acres
Category	5	6	7	8
Lot Size	3-5 acres	5-10 acres	10-20 acres	
Category	9	10	11	

To reflect the belief that half-bathrooms (i.e., those without a shower or bath-tub) are not valued by home-buyers nearly as highly as full bathrooms, the variable Bath records half-bathrooms with the value 0.1.

Table of Contents

Introduction

In this report, I will give an analysis of the information on the sale of 76 single-family homes in Dublin during 2005 and also the recommendation of single-family home sale price ($\frac{Price\ in\ Thousands\ of\ eur}{100{,}000}$) model.

Exploratory Data Analysis

Distribution of Sales Prices of the Houses

Here is the summary information of house price data:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.5   242.8   276.0   285.8   336.8   450.0
```

From the summary information, the range of minimum sale price is 155.5 and the maximum price is 450. The median of house price is 276 with means half of houses prices are greater than or equal to 276 and half are less. The middle 50% of house prices fall between 242.8 and 336.8. Figure 1 show that there are no potential outliers.

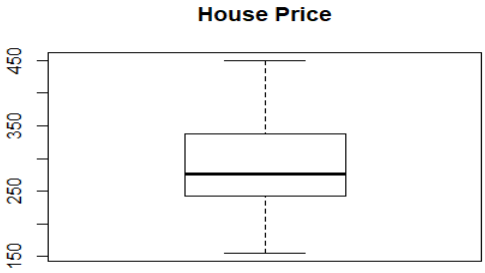


Figure 1 Boxplot of House Price

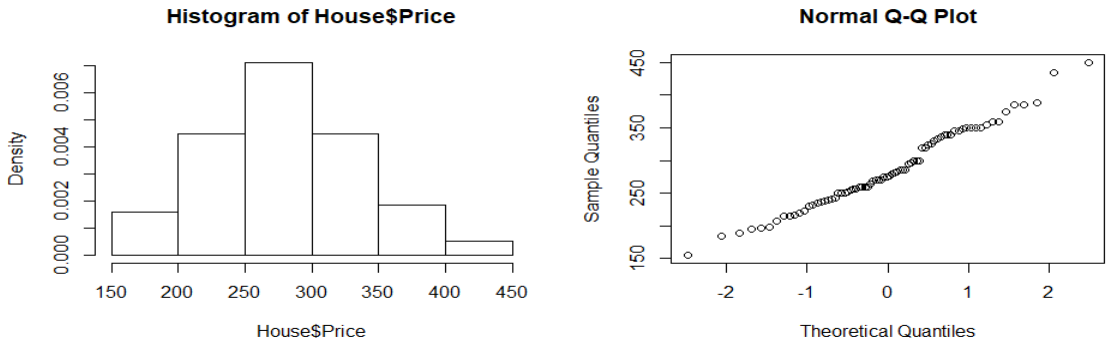


Figure 2 Distribution Plots of House Price

The histogram of distribution seems to be roughly symmetric bell-shaped because there are 2 out of 76 houses whose price between 400 and 450, but it is still can be considered as normal. The slope of normal Q-Q Plot also shows that the data is normally distributed. But there is a small gap in the middle, so far there is no significant pattern to be examined.

Here is the result of normality test:

```
##  Shapiro-Wilk normality test
##
## data:  House$Price
## W = 0.97981, p-value = 0.2671
```

H_0 = The population is normally distributed

The p-value is less than the chosen alpha level, then we failed to **reject the null hypothesis,** there is sufficient evidence that the data of house prices are **normally distributed**.

How Sales Prices vary based on the Categorical Variables

The sales prices vary with respect to some variables such as number of bedrooms, bathrooms, the garage size and also the school nearby the location of the houses.

```
##  Price          Size           Lot            Bath      Bed
##  Min.   :155.5  Min.   :1.440  Min.   : 1.000  1  : 2   2: 3
##  1st Qu.:242.8  1st Qu.:1.861  1st Qu.: 3.000  1.1: 5   3:43
##  Median :276.0  Median :1.966  Median : 4.000  2  :33   4:24
##  Mean   :285.8  Mean   :1.970  Mean   : 3.987  2.1:16   5: 5
##  3rd Qu.:336.8  3rd Qu.:2.107  3rd Qu.: 5.000  3  :13   6: 1
##  Max.   :450.0  Max.   :2.896  Max.   :11.000  3.1: 7

##       Year        Garage      School
##  Min.   :1905   0:11    Alex     : 3
##  1st Qu.:1958   1:13    High     :12
##  Median :1970   2:50    NotreDame:14
##  Mean   :1969   3: 2    StLouis  :15
##  3rd Qu.:1980           StMarys  :26
##  Max.   :2005           Stratford: 6
```

From the summary, it shows the number of houses related to each categorical variable. Here is the table of highest and lowest number of houses with a brief description.

Categorical Variable	Highest Number	Lowest Number
Bath	33 with 2 bathrooms	2 with 1 bathroom
Bed	43 with 3 bedrooms	1 with 6 bedrooms
Garage	50 with garage for 2 cars	3 with garage for 3 cars
School	26 near StMarys School	3 near Alex School

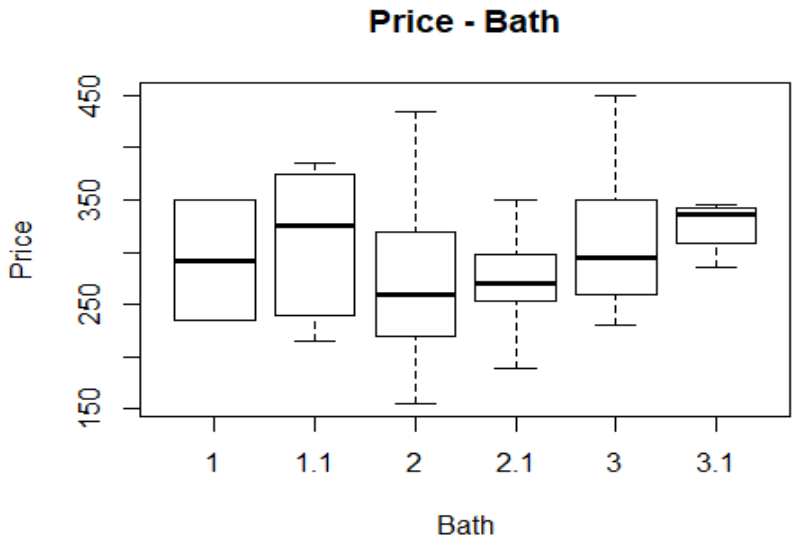


Figure 3 Sales Price respect to the number of Bathrooms

Based on the plot, 2 bathrooms has a wider range of house prices which also has a highest variance since it has 33 observations, otherwise 3 and a half bathrooms give least range of house prices whose median that seems to make shape of distribution a bit skew.

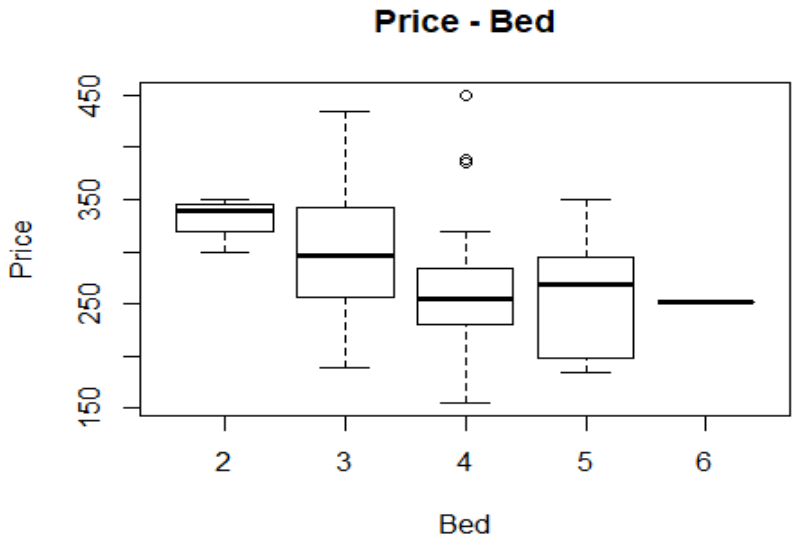


Figure 4 Sales Price respect to the number of Bedrooms

In this boxplot, 3 Bedrooms give wide range of prices which also has a highest variance since it has 43 observations, while 6 Bedrooms give the least one because it only has one data. Making the plot less informative. Other than that, there are 2 outliers within 4 Bedrooms boxplot. They are observation number 1,2 and 3 with the prices more than 350.

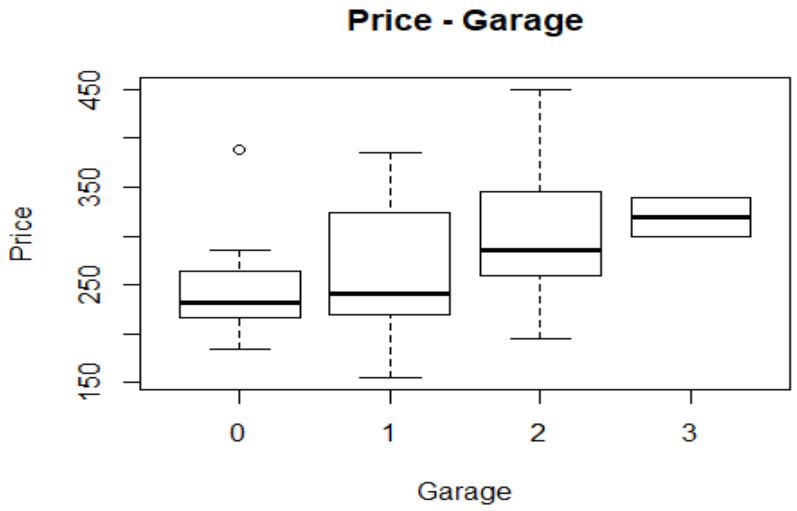


Figure 5 Sales Price respect to the Garage Size

From figure 3, it shows that as more spaces of cars (garage size) in the house, then the higher the house price. Garage for 2 cars has the wider range of prices which also has a highest variance since it has 50 observations, while Garage for 3 cars has the least range since it only has 2 observations. An outlier is spotted in the Garage 0 which is unusual for a house to be so expensive and have no space for any car.

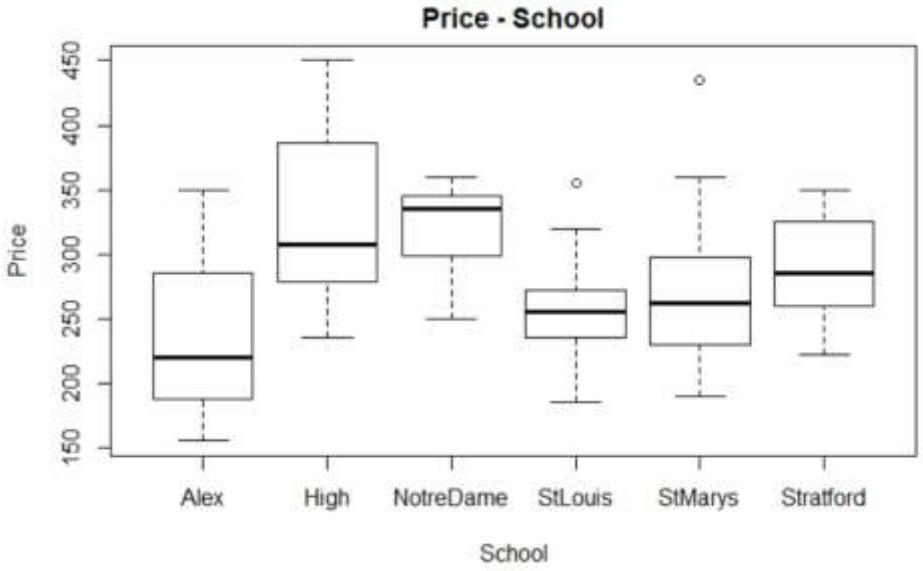


Figure 6 Sales Price respect to the School nearby

In this plot, the High School has wider range of house prices. NotreDame School has the highest median, while Alex School has the lowest median among the schools. There are 2 outliers spotted within StLouis School and StMarys School. Despite of the outlier, the boxplot of StMarys seems to be well distributed (median is located between Q1 and Q3) since it has sufficient number of observations.

How Sales Prices vary based on the Numerical Variables

Here is the summary with the highlight numerical variables:

```
##  Price          Size            Lot             Bath    Bed
##  Min.   :155.5  Min.   :1.440   Min.   : 1.000  1  : 2  2: 3
##  1st Qu.:242.8  1st Qu.:1.861   1st Qu.: 3.000  1.1: 5  3:43
##  Median :276.0  Median :1.966   Median : 4.000  2  :33  4:24
##  Mean   :285.8  Mean   :1.970   Mean   : 3.987  2.1:16  5: 5
##  3rd Qu.:336.8  3rd Qu.:2.107   3rd Qu.: 5.000  3  :13  6: 1
##  Max.   :450.0  Max.   :2.896   Max.   :11.000  3.1: 7

##       Year        Garage      School
##  Min.   :1905   0:11     Alex     : 3
##  1st Qu.:1958   1:13     High     :12
##  Median :1970   2:50     NotreDame:14
##  Mean   :1969   3: 2     StLouis  :15
##  3rd Qu.:1980            StMarys  :26
##  Max.   :2005            Stratford: 6
```

Based on the summary, for size variable the range is 1.440 and 2.896 with the close value between median and mean. The middle 50% of observations falls between size 1.861 and 2.107.

Lot summary shows the range of lot is from 1.000 and 11.000 with the close value between median and mean. The middle 50% of observations falls between lot 3.000 and 5.000.

Year summary shows that the oldest house is built in 1905 and the newest on 2005 with the close time between median and mean. 50% of houses were built during 1958 and 1980.

```
##              Price      Size       Lot        Year
## Price    1.0000000 0.20143783 0.24423228  0.15412476
## Size     0.2014378 1.00000000 0.04079199  0.17656934
## Lot      0.2442323 0.04079199 1.00000000 -0.03933975
## Year     0.1541248 0.17656934 -0.03933975 1.00000000
```

Correlation coefficient ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related. Based on the result, the correlation between house price and each variable seems to be **mildly** correlated because they are less than 0.30.

As r is positive, it means that as either Size, Lot or Year gets larger the Price gets greater. If r is negative it means that as one gets larger, the other gets smaller (often called an "inverse" correlation and in this case happened to Lot and Year variables).

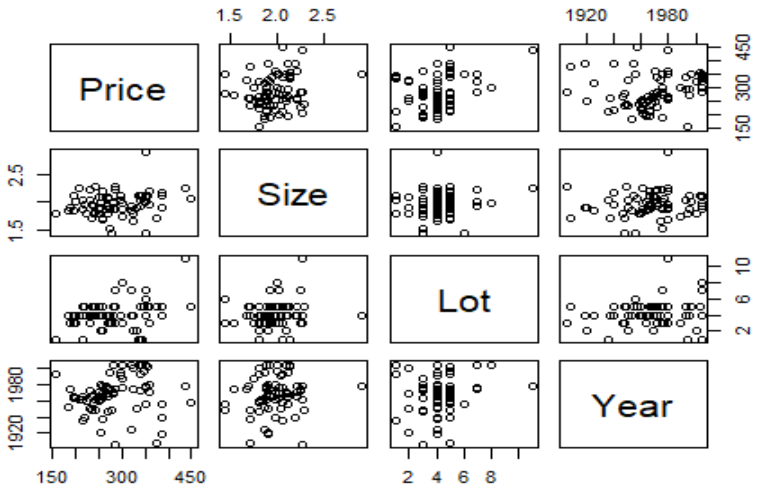


Figure 7 Pairs Plot of House Prices and Numerical Variables

A pairs plot allows us to see both distribution of single variables and relationships between two variables. In figure 7, there is no distinct pattern of relation between variables. The relationship between Lot and other variables seems to make parallel columns, those happened since the Lot is set as categories by the realtors but still there is no strong relationship within the pairs. Here are the plots in figure 8 with the closer look.

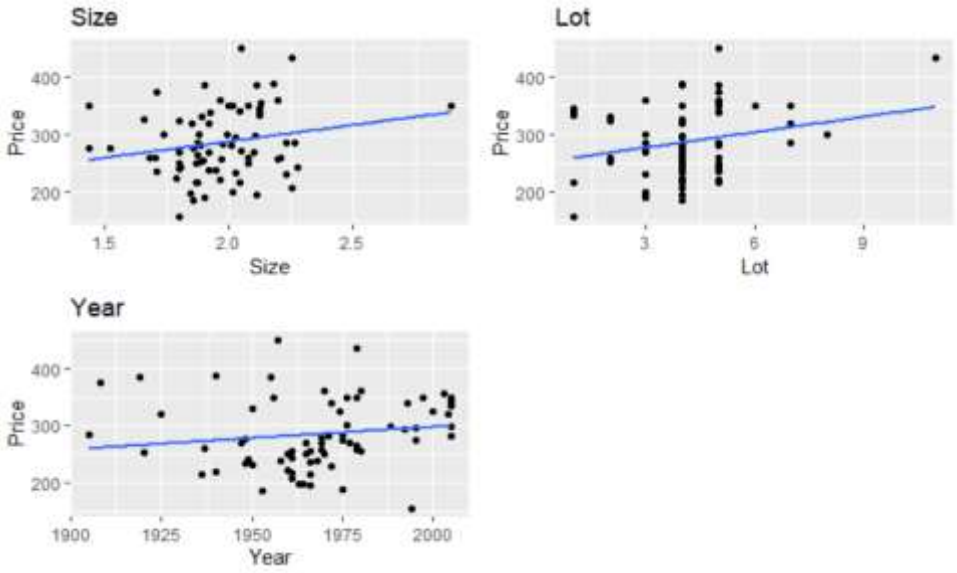


Figure 8 Size, Lot and Year relationship with Price

Regression Model

Multiple Linear Regression Model

$$\text{Modelprice} = \beta_0 + \beta_1\text{Size} + \beta_2\text{Lot} + \beta_3\text{Bath1.1} + \beta_4\text{Bath2} + \beta_5\text{Bath2.1} + \beta_6\text{Bath3} + \beta_7\text{Bath3.1} + \beta_8\text{Bed3} + \beta_9\text{Bed4} + \beta_{10}\text{Bed5} + \beta_{11}\text{Bed6} + \beta_{12}\text{Year} + \beta_{13}\text{Garage1} + \beta_{14}\text{Garage2} + \beta_{15}\text{Garage3} + \beta_{16}\text{SchoolHigh} + \beta_{17}\text{SchoolNotreDame} + \beta_{18}\text{SchoolStLouis} + \beta_{19}\text{SchoolStMarys} + \beta_{20}\text{SchoolStratford} + \varepsilon$$

```
## Call:
## lm(formula = Price ~ ., data = House)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -87.601 -21.429   0.173  24.248  72.58    1
##
## Coefficients:
##                      Estimate    Std. Error    t value    Pr(>|t|)
## (Intercept)          -884.3531   661.7589      -1.336     0.18693
## Size                 59.4503     28.9813       2.051      0.04501 *
## Lot                  11.7701     3.7842        3.110      0.00296 **
## Bath1.1              135.8983    49.1990       2.762      0.00779 **
## Bath2                73.9317     47.8636       1.545      0.12817
## Bath2.1              76.9433     48.1208       1.599      0.11556
## Bath3                98.0694     50.4663       1.943      0.05711 .
## Bath3.1              85.8037     54.3074       1.580      0.11985
## Bed3                 -228.1052   70.6732       -3.228     0.00211 **
## Bed4                 -238.2609   72.4883       -3.287     0.00177 **
## Bed5                 -237.6155   76.4733       -3.107     0.00299 **
## Bed6                 -255.0211   88.0955       -2.895     0.00543 **
## Year                 0.5567      0.3384        1.645      0.10565
## Garage1              -10.9191    22.4871       -0.486     0.62920
## Garage2              18.2435     18.2212       1.001      0.32111
## Garage3              -209.9038   80.7191       -2.600     0.01193 *
## SchoolHigh           113.2774    36.9154       3.069      0.00334 **
## SchoolNotreDame      80.9317     35.6893       2.268      0.02730 *
## SchoolStLouis        9.0367      37.3439       0.242      0.80969
## SchoolStMarys        27.3408     35.8760       0.762      0.44926
## SchoolStratford      31.9254     40.9171       0.780      0.43859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.13 on 55 degrees of freedom
## Multiple R-squared:  0.6425, Adjusted R-squared:  0.5125
## F-statistic: 4.942 on 20 and 55 DF,  p-value: 1.265e-06
```

The Intercept

The estimate of the intercept term β_0 in this model is -884.3531. In this case, the intercept is not interpretable. The intercept starts to make sense when the variables are in the range of possible values, for instance Size variable has value in range of 1.440 and 2.896 or Year has value in range of 1905 and 2005. Those values are far from 0. It is unlikely that a zero is a valid indicator for the rest of variables.

The β_{size}

The estimate of the β_{size} the parameter associated with floor size in this model is 59.4503, means that if β_{size} increasing the size by one unit (thousands of square feet), will result in an extra 59.4503 of expected value of house price.

The $\beta_{Bath1.1}$

The estimate of the $\beta_{Bath1.1}$ the parameter associated with one and a half bathrooms in this model is 135.8983, means that the expected value of house price increasing by 135.8983 while the house is having one and a half bathrooms.

The Effect of Predictor Variable Bed

There are 4 categories in variable Bed, and one of it (in this case is Bed 2) by default is calculated and 'contained' within the value of intercept. For instance, if the house has 2 bedrooms, the model will have the intercept to 'represent' the value of β_{Bed2}.

For the rest of parameters of bed such as β_{Bed3}, β_{Bed4}, β_{Bed5} and β_{Bed6}, if the house has any of 3, 4, 5 or 6 bedrooms, then the house prices will drop by 228.1052, 238.2609, 237.6155, or 255.0211 respectively because the value of estimators are negative respect to $\beta_{Bed2.}$

Significant Predictor Variables

Here is the list of predictor variables that are significantly contributing to the expected value of the house prices:

1. Size
2. Lot
3. Bath
4. Bed
5. Garage
6. School

Values those lead to Largest and Lowest Expected Value of the House Prices

1. Size = 59.4503
2. Lot = 11.7701
3. Bath = 135.8983 by $\beta_{Bath1.1}$ (largest) and β_{Bath1} which within the intercept (lowest)
4. Bed = β_{Bed2} which within the intercept since the rest give negative value (largest) and β_{Bed6} with the value -255.0211 (lowest)

5. Garage = 18.2435 by $\beta_{Garage2}$ (largest) and $\beta_{Garage3}$ with -209.9038 (lowest)
6. School = 113.2774 by $\beta_{SchoolHigh}$ (largest) and $\beta_{SchoolAlex}$ which within the intercept (lowest)
7. Year = 0.5567

Overall, by comparing all the parameters within the variables, the value of $\beta_{Bath1.1}$ lead to the largest and the value of β_{Bed6} lead to the lowest expected value of the house prices.

Residuals of the Expected Value of the House Prices

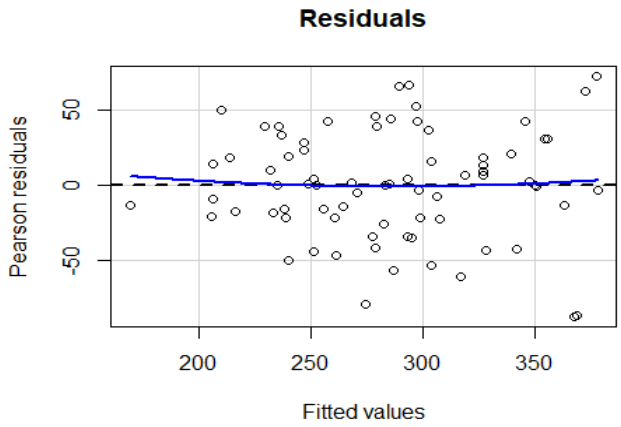


Figure 9 Residuals

Plot shows that the residuals appear to behave randomly; it suggests that the model fits the data well. From the summary, the residual standard error which is the average amount that the price will deviate from the intercept -884.3531 is 42.13 we can say that the percentage of error is 4.7%. In my opinion, we need to do further analysis to make sure that this is a good model of the expected value of the house prices even though the residuals seems to be good.

Adjusted R-squared value

Typical values range from 0.1-0.9. In this model, we arrived at an adjusted R-squared value of 0.5125 indicating that 51.25% of the variation in price is explained by the model.

F-Statistic

F-Statistic is a test statistic to check if the data conforms to a regression model, based on the output in the summary of the regression model here is the interpretation:

H_0 = all betas equal to 0s

H_A = at least one of betas is not equal to 0

The value of F-Statistic is 4.942 on 20 and 55 DF; and p-value is 1.265e-06 which is less than 0.05, so we **reject** the null hypothesis and conclude that at least one of the betas is non-zero.

ANOVA

Type 1 Anova

```
## Analysis of Variance Table
## Response: Price
##             Df  Sum Sq  Mean Sq F value    Pr(>F)
## Size         1   11078  11077.7  6.2426  0.015489 *
## Lot          1   15232  15232.5  8.5839  0.004929 **
## Bath         5   36824   7364.7  4.1502  0.002861 **
## Bed          4   25502   6375.4  3.5927  0.011310 *
## Year         1     554    554.4  0.3124  0.578474
## Garage       3   16101   5367.1  3.0245  0.037179 *
## School       5   70112  14022.4  7.9020 1.153e-05 ***
## Residuals   55   97599   1774.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We use F-Statistic in anova type 1 and the value of F-Statistic is calculated for each variable and marked as significant by the signif. codes. The hypotheses are:

H_0 = beta is equal to 0
H_A = beta is not equal to 0

The only variable that is not marked as significant here which has p-value more than 0.05 is 'Year', so for this variable, we failed to reject that the null hypothesis for variable Year is equal to 0. However, the rest of variables have p-value less than 0.05 which we may reject the null hypotheses and conclude that their betas are not equal to 0.

Then, type 1 anova table recommend that we should remove variable **'Year'** from the regression analysis.

Type 2 Anova

Within this step, type 2 Anova is generated

```
## Analysis of Variance Table
## Model 1: Price ~ Size + Lot + Bath + Bed + Year + Garage + School
## Model 2: Price ~ Size + Lot + Bath + Bed + Garage + School
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     55  97599
## 2     56 102402 -1   -4802.6 2.7064 0.1057
```

F-Statistic is used and the hypotheses:

H_0 = β_{Year} is equal 0

$H_A = \beta_{Year}$ is not equal to 0

The probability is more than 0.05, then we failed to reject null hypothesis and conclude that in this case removing variable 'Year' should be done to improve the model.

Diagnostics

Linearity

Here is the checking of linearity assumption. I plotted the added variable plots to check the effect of a particular predictor on house price while holding all other predictor. As the slope of the fitted line is different from zero, the variables have a significant impact to the model.

From the plots, those variables are: Size, Lot, Bath (all), Bed (all), Garage3, SchoolHigh, SchoolNotreDame. Meanwhile, the variables with the slope closer to zero are Garage1, Garage2, SchoolStLouis, SchoolStMarys and SchoolStratford. They do not have a significant impact to the model.

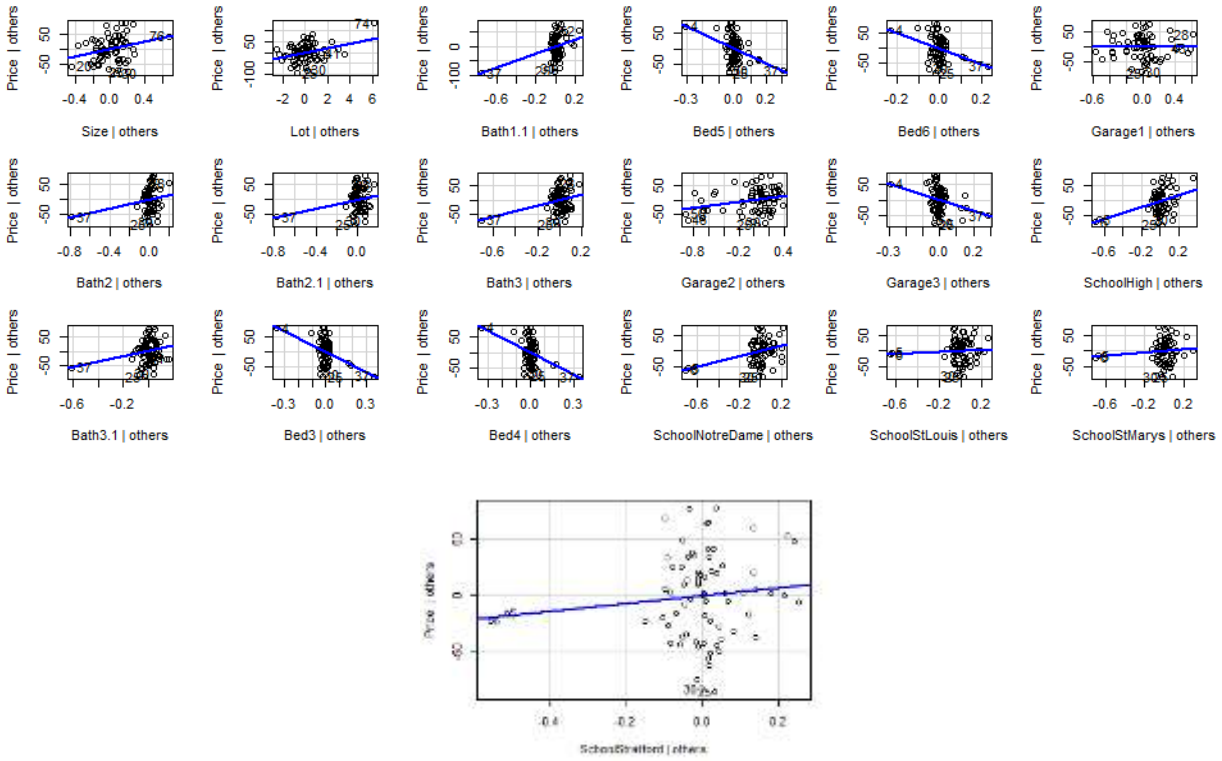


Figure 10 Added Variable Plots

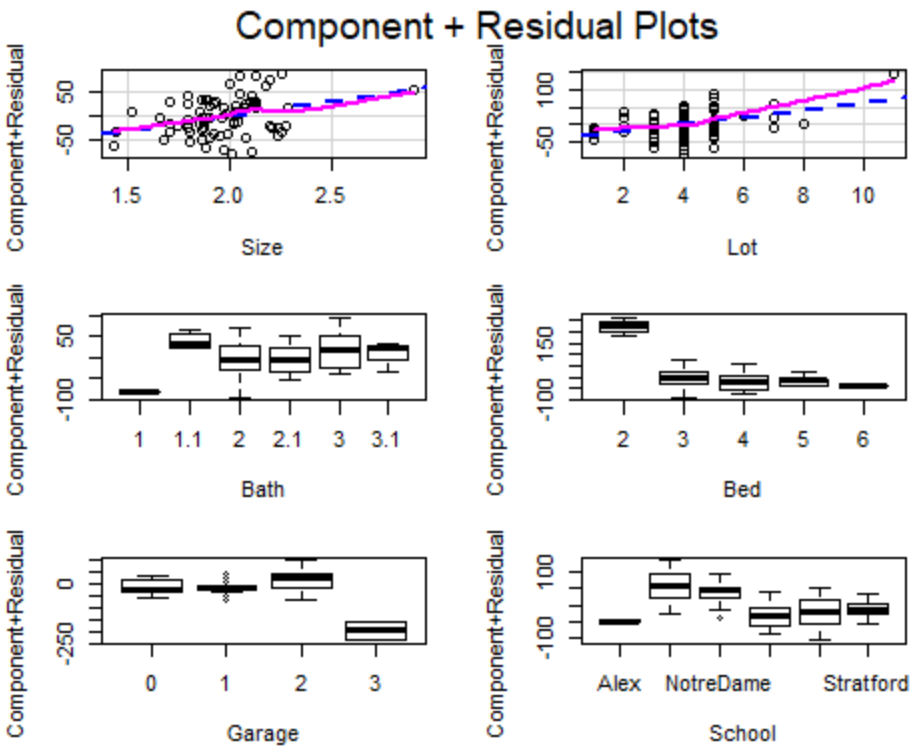


Figure 11 Component and Residuals Plots

Both component residual plots of size and lot related to house price are approximately linear as the dashed and the pink line do not differ dramatically. To be noted, there is a potential outlier point in the Lot plot which drags pink dashed line upright.

The effect non-linearity would have on the regression model is the value for one observation is likely to be influenced by the value of another observation. To improve the model in the presence of non-linearity, we can use transformation to change the linear relationship between variables or splines to smoothly interpolate between fixed points.

Random/i.i.d.

Here is the checking of the random/i.i.d sample assumption by computing the Durbin Watson test statistic.

```
##  lag           Autocorrelation       D-W Statistic   p-value
##    1           0.2316982             1.511734        0.002
##  Alternative hypothesis: rho != 0
```

H_0 = rho is equal 0
H_A = rho is not equal to 0

The result of Durbin Watson test statistic is 1.511 and the p-value is 0.002 so the hypothesis of no autocorrelation is rejected, and the observations cannot be classed as independent.

Two common violations of the random/i.i.d. sample assumption: The samples are not independent, and they do not have the same variance. So, in my opinion, maybe size and some of other variables in the model may have a dependency. For example, Bedrooms, Bathrooms and Garage related to the Size or Lot. It makes sense for instance that the more bedrooms in the house, the more spaces that the house has.

Having dependent samples would skew the data on the regression model, value for one observation is likely to be influenced by the value of another observation, they will affect the model and rises the risk that all of the results will be wrong.

To improve the model in the presence of dependent samples, we can use Mixed Effect Models for Repeated Measurements or Time Series Analysis for Correlation in the errors.

Multicollinearity

Here is checking the collinearity assumption by interpreting the correlation and variance inflation factors.

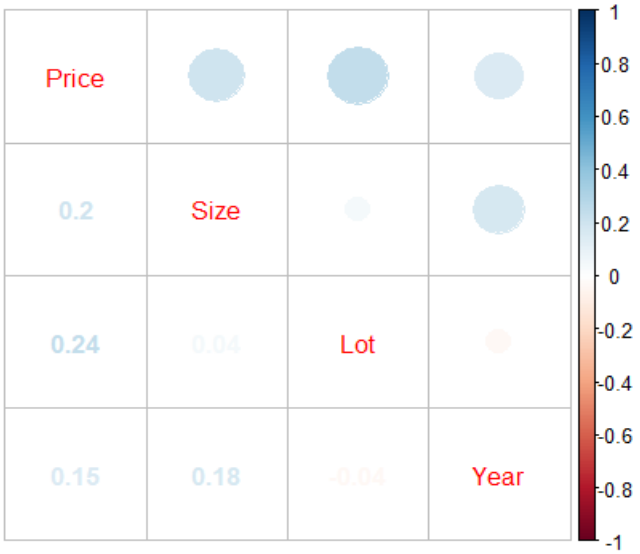


Figure 12 Correlations Plots

The correlation between variables with their correlation are completely mild and there is no multicollinearity within the model. Indicating that it is unlikely we will have a problem with a regression including these predictor variables.

```
##              GVIF       Df     GVIF^(1/(2*Df))
## Size     1.599498       1          1.264713
## Lot      1.586836       1          1.259697
## Bath     8.302874       5          1.235728
## Bed     17.585637       4          1.431017
## Garage  15.567849       3          1.580173
## School   5.342986       5          1.182438
```

The variance inflation factors are approximately around 1-1.5, indicating that we don't have a multicollinearity problem with a regression.

If there is a strong correlation (multicollinearity) between the regressors, then the parameters become unstable. The estimate of parameters will depend strongly on the other predictors that are included in the model.

To improve the model in the presence of multicollinearity we may remove highly correlated predictors from the model or by using Partial Least Squares Regression (PLS), Principal Components Analysis, or Ridge Regression those will cut the number of predictors to a smaller set of uncorrelated components.

Zero Conditional Mean and Homoscedasticity

Here is checking the zero conditional mean and homoscedasticity assumptions. In all plots, there is a mild funnel; means there is a small non-constant variance within the model.

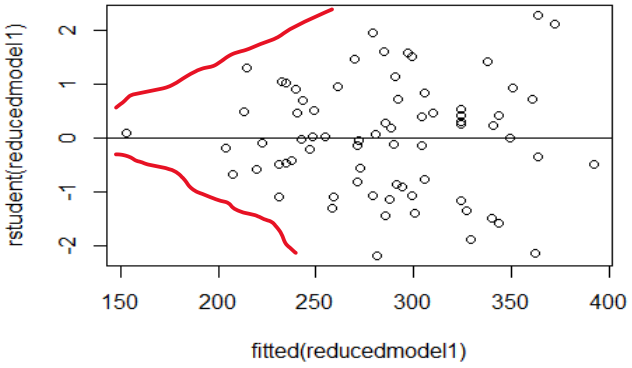


Figure 13 The Studentized Residuals vs Fitted Values

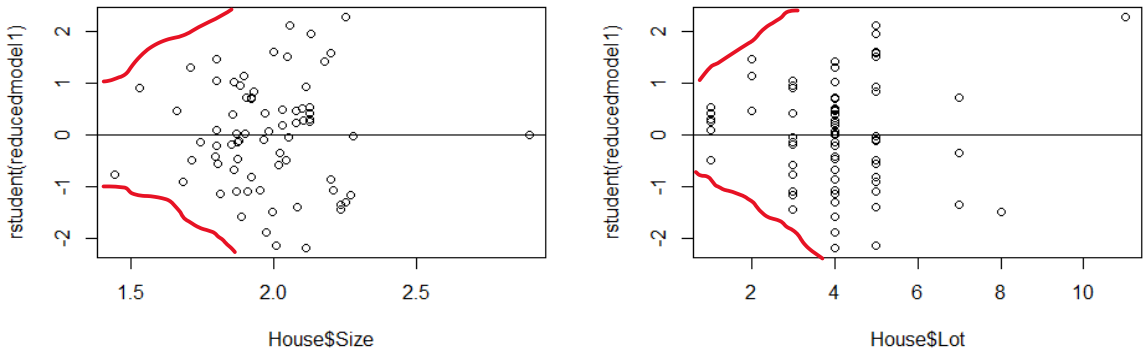


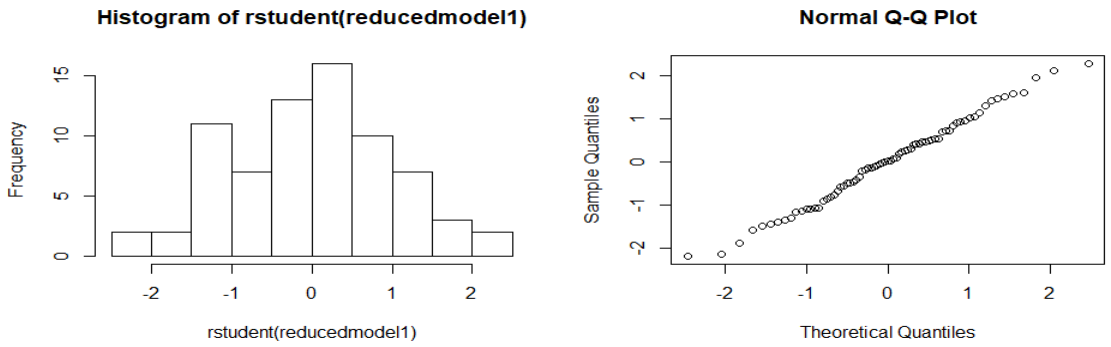
Figure 14 The Studentized Residuals vs Predictor Variables

Heteroscedasticity tends to produce p-values that are smaller than they should be. This effect occurs because heteroscedasticity increases the variance of the coefficient estimates.

To improve the model in the presence of heteroscedasticity we can use Weighted Least Squares for Non-Constant Variance (when the funnel shapes in residual vs fitted values plot indicate non-constant variance) or examine the outliers. Related to the house price modeling, since the zero conditional mean and homoscedasticity assumption need to be fully satisfied, there will be an outlier test within the next part of *Leverage, Influence and Outliers.*

Normality

Here is the checking of the normality assumption of the studentized residuals.



The histogram of studentized residuals seems to make a bell shape with a mildly stretch in the left (-1) area and the slope of normal Q-Q Plot of the studentized residuals also shows that the data is normally distributed.

The effect of non-normality on the regression model is the wrong result of critical values for t and F tests. To improve the model in the presence of non-normality we can do the transformations, interactions, or try to build a different model.

Leverage, Influence and Outliers

Leverage

A leverage point is the observation with an unusual X-value. It will affect the regression model summary but might have little effect on the estimates of the regression coefficients. However, the high leverage points have the potential to affect the fit of the model.

I made a code to define the hat values. There are 3 observations with the value equal to one. They are observation number 4, 35 and 37.

```
## hatvalues(reducedmodel1) warn
## 1          0.29606446   -
## 2          0.20929675   -
## 3          0.18703613   -
## 4          1.00000000   -
## 5          0.55012411   -
## 6          0.55012411   -
## 7          0.32949682   -
## 8          0.13616235   -
## 9          0.17921740   -
## 10         0.14299187   -
## 11         0.11055188   -
## 12         0.10897094   -
## 13         0.09688044   -
## 14         0.07872214   -
## 15         0.20753681   -
## 16         0.14894057   -
## 17         0.13916846   -
## 18         0.17262452   -
## 19         0.10644515   -
## 20         0.25717324   -
## 21         0.59085455   -
## 22         0.37053025   -
## 23         0.17709424   -
## 24         0.11730752   -
## 25         0.09446984   -
## 26         0.13381121   -
## 27         0.11688571   -
## 28         0.31795374   -
## 29         0.08421354   -
## 30         0.14432214   -
## 31         0.38096707   -
## 32         0.30827758   -
## 33         0.21006526   -
## 34         0.19924956   -
## 35         1.00000000   -
## 36         0.33252819   -
## 37         1.00000000   -
## 38         0.13092637   -
## 39         0.18552479   -
## 40         0.13676870   -
## 41         0.46000098   -
## 42         0.25003283   -
## 43         0.22898529   -
## 44         0.14529767   -
## 45         0.14651273   -
## 46         0.21707574   -
## 47         0.59085455   -
## 48         0.11081630   -
## 49         0.27022875   -
## 50         0.31255832   -
## 51         0.35537581   -
## 52         0.32401693   -
## 53         0.15653789   -
## 54         0.37567984   -
## 55         0.15909315   -
## 56         0.19445483   -
## 57         0.23739985   -
## 58         0.17631073   -
## 59         0.16674051   -
## 60         0.16674051   -
## 61         0.16674051   -
## 62         0.16674051   -
## 63         0.16674051   -
## 64         0.32640617   -
## 65         0.14252070   -
## 66         0.20780412   -
## 67         0.11690853   -
## 68         0.14141256   -
## 69         0.25115682   -
## 70         0.13663613   -
## 71         0.23860758   -
## 72         0.27315568   -
## 73         0.41946187   -
## 74         0.41584296   -
## 75         0.13136315   -
## 76         0.61451059   -
```

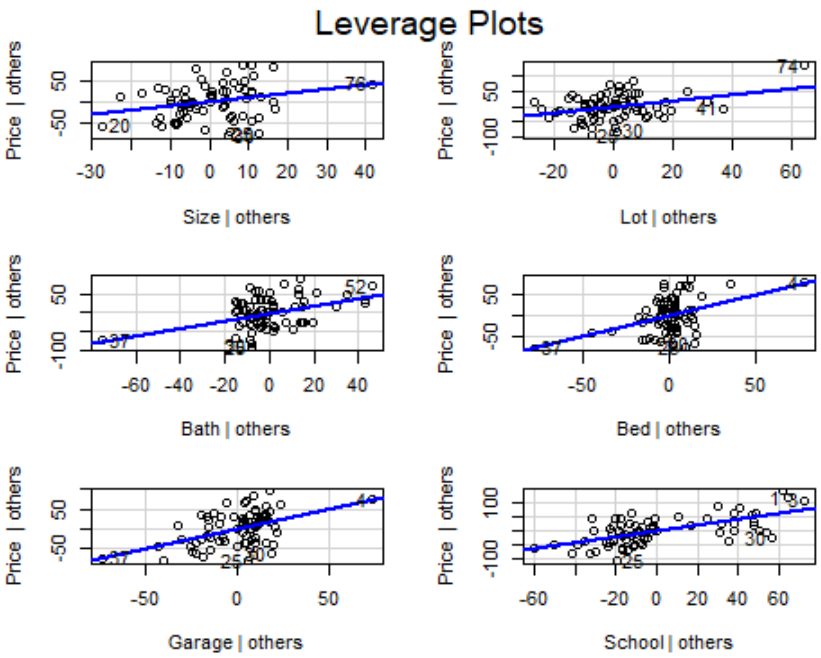


Figure 15 The Leverage Plots

In Bed and Garage leverage plots, we can see that the observation number 4 and 37 are trying to drag the blue lines. Not only there, observation number 37 also appears in Bath leverage plot. In addition from Lot Leverage Plot, the observation number 74 also one of the potential leverage points. Below is the residuals vs leverage plot, there was a warning message while loading the plot: "not plotting observations with leverage one: **# 4, 35, 37**"

Figure 16 The Residuals vs Leverage Points

When the observations are outside of the Cook's distance (meaning they have high Cook's distance scores), the cases are influential to the regression results. The regression results will be altered if we exclude those cases.

Influence

An influential point may either be an outlier or have large leverage, or both, but it will tend to have at least one of those properties. It has a noticeable impact on the model coefficients: it 'drags' the regression model in its direction.

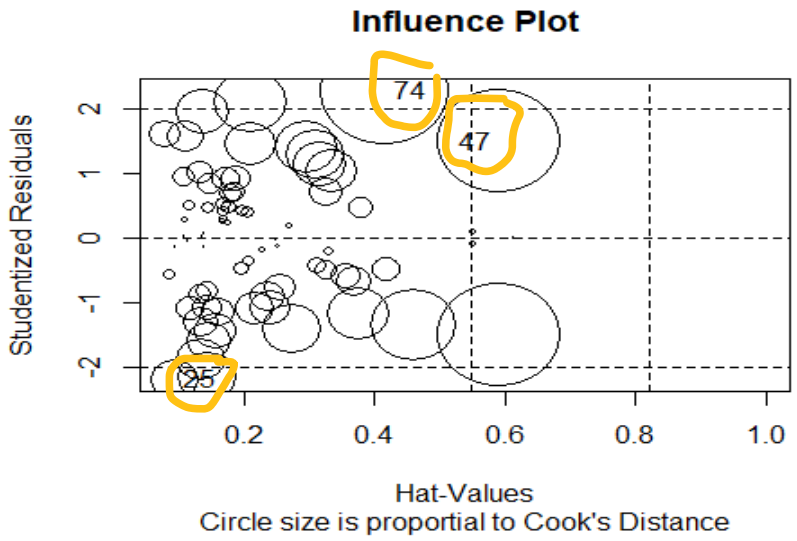


Figure 17 The Influence Plot

In the influence plot, we can see that there are 3 observations that might influence the model; they are observations number 25, 47 and 74. The summary of influence points includes observation number 4 and 35 (as I mentioned before) as influencer points.

##	StudRes	Hat	CookD
## **4**	NaN	1.00000000	NaN
## 25	-2.183451	0.09446984	0.02330078
## **35**	NaN	1.00000000	NaN
## 47	1.499325	0.59085455	0.15877875
## 74	2.276139	0.41584296	0.17159204

Outliers

An outlier is a data point that differs significantly from other observations. An outlier can cause serious problems in statistical analyses. It might affect the estimation of the regression coefficients. Since it may be due to variability in the measurement, typo or it may indicate experimental error.

To correct the outliers, we can do:
1. Drop the outlier observations.
2. Cap the outlier data.
3. Assign a new value.
4. Try a transformation.

Here is the outlier test for the model, it is Bonferroni p-values for testing each observation in turn to be a mean-shift outlier, based Studentized residuals in linear (t-tests), generalized linear models (normal tests), and linear mixed models. If there is p-value of residual that is less than 0.05, it will be considered as an outlier. In this case, there is no residual that can be considered as outlier.

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 74 2.276139          0.026753           NA
```

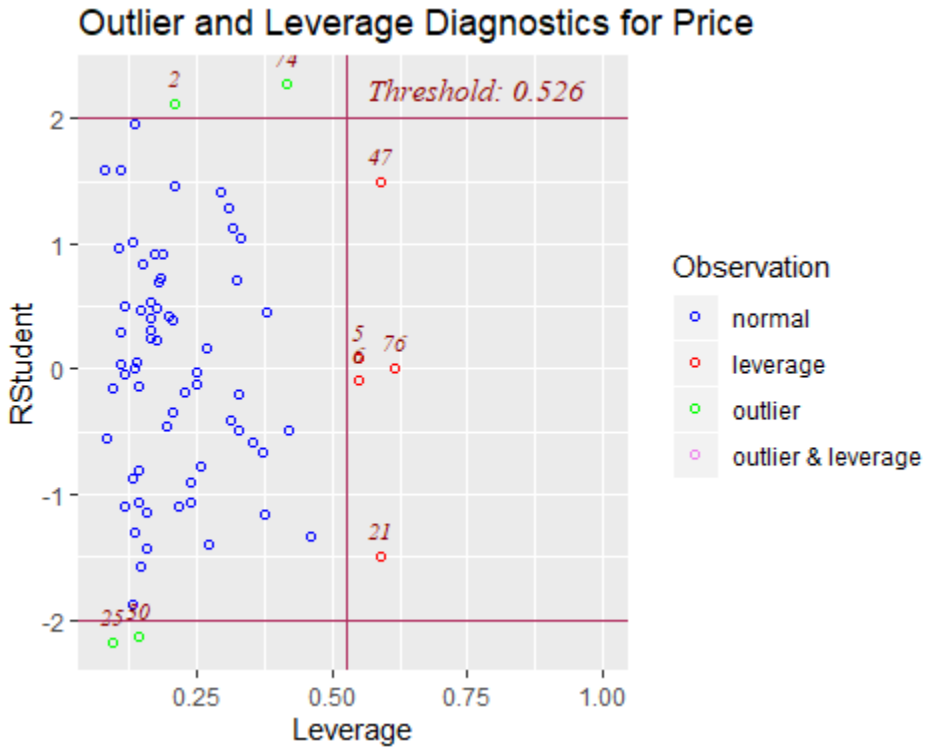


Figure 18 The Outlier and Leverage Diagnostics Plot

The plot shows 4 outlier observations, but because the observations are still in the range of -3 and 3, they are still can be considered as normal and I will not remove them for now.

Next, I want to examine their details further by comparing the similarity.

Here are the points whose value more than 2 but less than 3.

```
House[74,]

##      Price  Size   Lot    Bath   Bed    Year   Garage       School
## 74   435    2.253  11     2      3      1979   2            StMarys

House[2,]

##      Price  Size   Lot    Bath   Bed    Year   Garage       School
## 2    450    2.054  5      3      4      1957   2            High
```

Both observations have the price over 400, Garage Size 2 and also Size those over 2.000. It means that the price can be classified as more expensive while they have greater size including spaces for 2 cars.

Then, here are the points whose value more than -2 but less than -3.

```
House[25,]

##      Price  Size   Lot    Bath   Bed    Year   Garage       School
## 25   195    2.112  4      2      3      1966   2            StMarys

House[30,]

##      Price  Size   Lot    Bath   Bed    Year   Garage       School
## 30   279.9  2.01   5      2      3      1969   2            High
```

Both observations have the Garage Size 2, 2 Bathrooms, 3 Bedrooms and also Size those over 2.000. It means that the houses have greater size which makes sense to support the spaces needed by the number of Bathrooms, Bedrooms and the Garage in it.

Expected Value, CI and PI

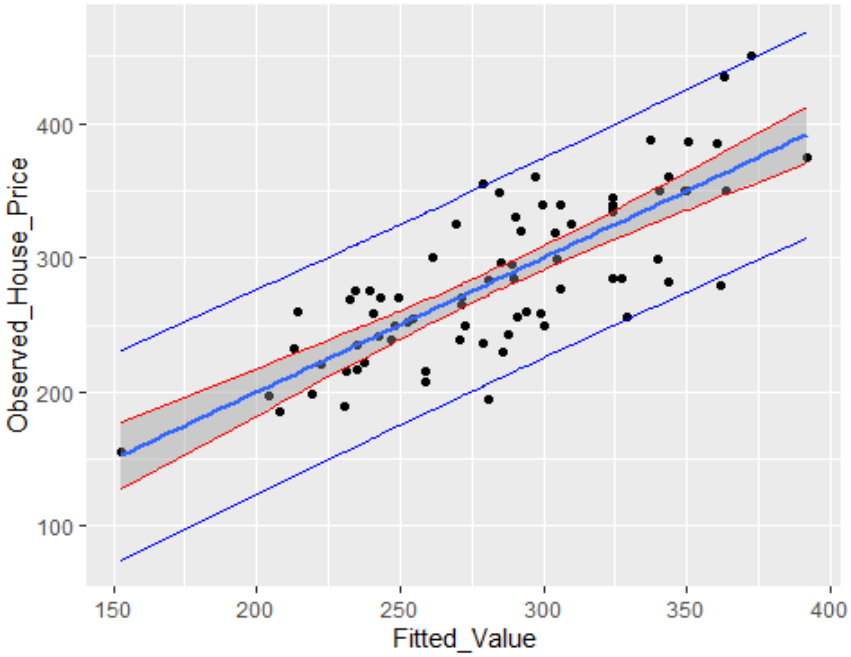


Figure 19 The Model Plot

With level of 95% significant, we can say that if we would repeat our sampling process infinitely, 95% of the constructed confidence intervals (within the red lines) would contain the true house prices mean and 95% of the constructed prediction intervals (within the blue lines) would contain the new observation. From figure 19, we can say that the model can be considered providing the good estimate of the house prices. But, this might not be the perfect model; there must be a room for any improvement.

For the future studies, I recommend to either define the rooms (bedrooms, bathrooms and garage) with the metric size e.g. in meters and not including lot/size altogether; or just including lot or size while excluding the rooms. I am expecting within that recommendation, the model will be simpler and we can avoid any unnecessary correlation or maybe sample dependency those might giving significant effect.