Predictive Analytics Assignment Report

Raj Shekhar

18200277

Data & Computational Science

There are total 8 variables in the House dataset.
There are 4 categorical variables: Bath, Bed, Garage and School as they have five or less than five unique value. The rest Price, Size, Lot and Year are numeric variables.

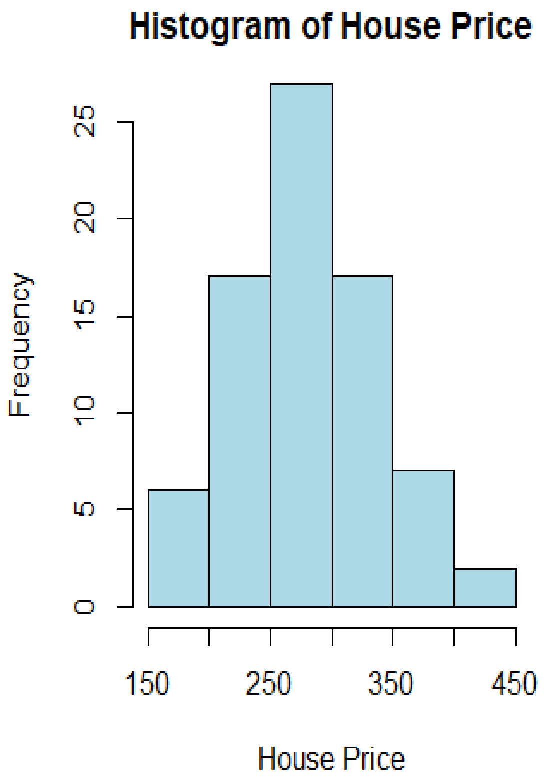
Exploratory Data Analysis:

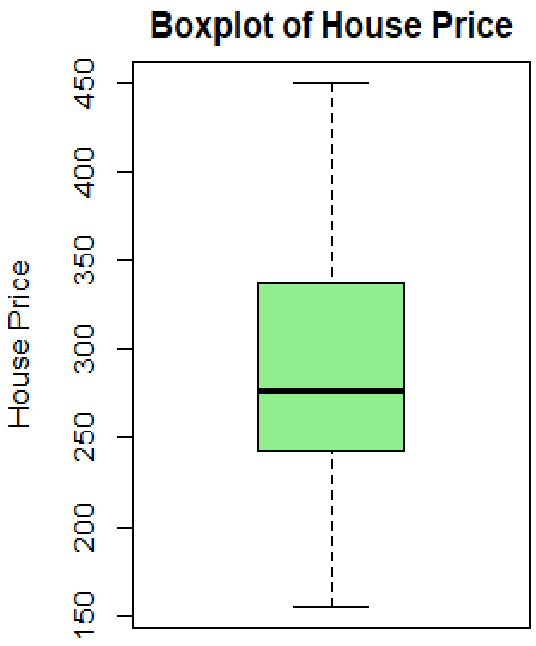
It is an approach of analyzing data sets to summarize their main characteristics, often with visual methods. EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task using the provided data.

```
# Loading Dublin House Dataset

House <- read.csv(choose.files(),header = TRUE,row.names = NULL)
```

1. Using the boxplot, histogram and summary. Describe the distribution of sale price of the houses.

- Histogram and Box plot of House Price:





- **Summary Statistic of all variables:**

- Categorical variables:

1. **Bath** : The categorical variable shows number of bathrooms (1, 1.1, 2, 2.1, 3 & 3.1)
2. **Bed** : Shows number of bedrooms (between 2 and 6)
3. **Garage** :Shows the garage size (0, 1, 2, & 3)
4. **School** : Shows school as High, Alexandra, Stratford, St. Mary's, NotreDame and St Louis.

- Numerical Variables:

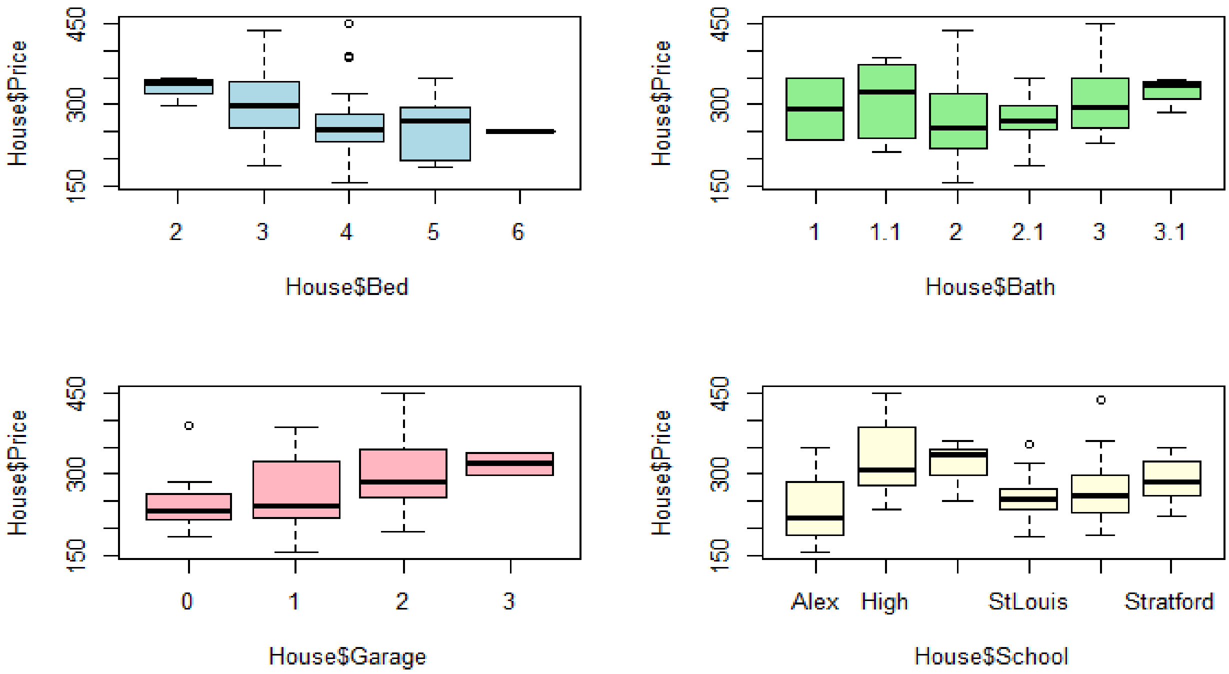
1. **Price**: This is the target variable of the dataset where the model is run against it to predict the overall house price. Minimum observed is:155.5, Maximum is 450.The 25% house have price below 242.8 whereas, 25% house have price above 336.8.
   About 50% of house have price approximately below 275 and 50% above that. Mean is 285.8 and Median is 276.0. Histogram is right skewed.
2. **Size**: Represents floor size (thousands of square feet). Minimum size observed is:1.440, Maximum is 2.896. 25% house have size below 1.861 whereas, 25% house have size above 2.107. Mean is 1.970 and Median is 1.966. Mean is almost equal to median.
3. **Lot** : Represents lot size category . Minimum weight observed is:1, Maximum is 11. 25% house are below Lot=3 whereas, 25% house are above Lot=5. Mean is 3.987 and Median is 4. As mean is almost equal to median.
4. **Year** : Represents year the house was built.. Minimum year observed is:1905 , Maximum is 2005. 25% houses were built before 1958 whereas, 25% house were built after 1980. Mean is 1969 and Median is 1970. As mean lesser than the median.

```
summary(House)
```

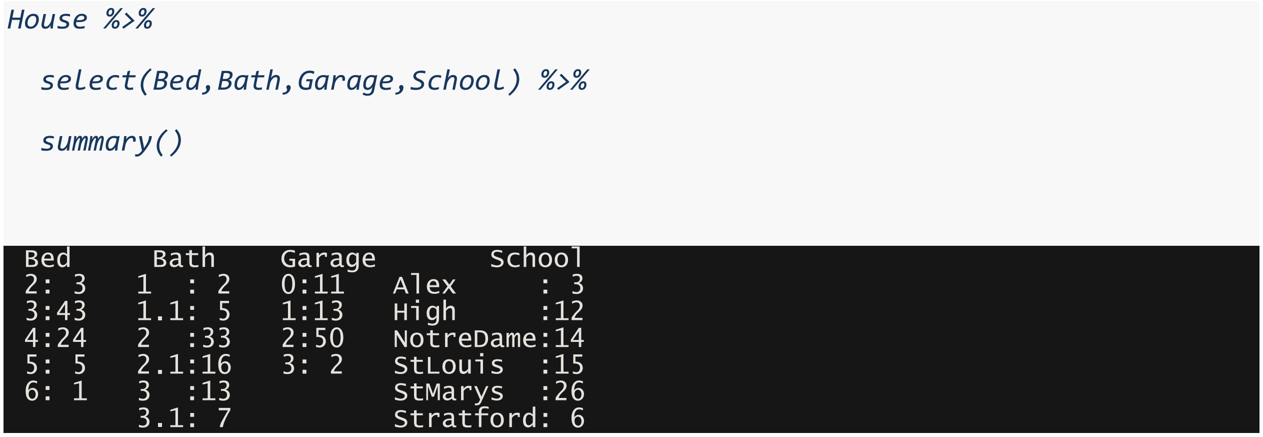
Price	Size	Lot	Bath	Bed	Year	Garage	School
Min. :155.5	Min. :1.440	Min. : 1.000	Min. :1.000	Min. :2.000	Min. :1905	Min. :0.000	Alex : 3
1st Qu.:242.8	1st Qu.:1.861	1st Qu.: 3.000	1st Qu.:2.000	1st Qu.:3.000	1st Qu.:1958	1st Qu.:1.000	High :12
Median :276.0	Median :1.966	Median : 4.000	Median :2.000	Median :3.000	Median :1970	Median :2.000	NotreDame:14
Mean :285.8	Mean :1.970	Mean : 3.987	Mean :2.208	Mean :3.447	Mean :1969	Mean :1.566	StLouis :15
3rd Qu.:336.8	3rd Qu.:2.107	3rd Qu.: 5.000	3rd Qu.:3.000	3rd Qu.:4.000	3rd Qu.:1980	3rd Qu.:2.000	StMarys :26
Max. :450.0	Max. :2.896	Max. :11.000	Max. :3.100	Max. :6.000	Max. :2005	Max. :3.000	Stratford: 6

2: Convert all the categorical variables to factors. Using the summary and a boxplot describe how sales prices vary with respect to the number of bedrooms, bathrooms, garage size and school.

- Box Plots of Categorial Variables



- Summary of 4 categorical variable



```
House %>%

  select(Bed,Bath,Garage,School) %>%

  summary()
```

```
 Bed      Bath     Garage        School
 2: 3    1  : 2   0:11    Alex     : 3
 3:43    1.1: 5   1:13    High     :12
 4:24    2  :33   2:50    NotreDame:14
 5: 5    2.1:16   3: 2    StLouis  :15
 6: 1    3  :13           StMarys  :26
         3.1: 7           Stratford: 6
```

Summary of each categorical variable with dependent variable price:

Boxplot of Price Vs Beds

There are 5 boxplots generated representing distribution of number of beds corresponding to price.

- For **Bed 2**, the mean is closer to 3rd quartile and the distribution is left skewed.House Sale price ranges between min of 299 and max of 350
- For **Bed 3**, the mean is almost at the center. House Sale price ranges between min of 189.5 and max of 435 and they have most number of houses.
- For **Bed 4**, the mean is closer to 1st quartile and the distribution is right skewed with few outliers. House Sale price ranges between min of 155.5 and max of 450
- For **Bed 5**, the distribution is closer to the 3rd quartile. House Sale price ranges between min of 185 and max of 349.5
- For **Bed 6**, there is one value and hence line is being plotted for it. It has only one house with value of 252.5
- Specific point to highlights is:
  - Outliers can be observed for 4 Bed House.
  - For Bed 6, the mean coincides with the 1st & 3rd quartile value
  - No Bed house shows right skewness

Boxplot of Price Vs Bath

There are 6 boxplots generated representing distribution of number of bathroom corresponding to overall house price.

- For **Bath 1**, It shows almost symmetric shape hence symmetric normal distribution. House price ranges between 235-350.
- For **Bath 1.1**, mean is closer to the 3rd quartile. Also, maximum is closer to the 3rd quartile than the minimum from 1st quartile. House price ranges between 215-385.5.
- For **Bath 2**, Minimum is very close to the first quartile, indicating that there are very less values in this region. Maximum is far from 3rd quartile hence the distribution is right skewed. House price ranges between 155.5 - 435 and has highest number of houses.
- For **Bath 2.1**, min value is almost at equal distance from 1st quartile as the distance of maximum from 3rd quartile. Mean is closer to1st quartile. The distribution is right skewed. House price ranges between 189.5 – 349.5.
- For **Bath 3**, mean is closer to the 1st quartile. House price ranges between 230 – 450.
- For **Bath 3.1**, mean is closer to the 3rd quartile. Also, the maximum is very close to the 3rd quartile as compared to the distance of minimum from 1st quartile. House price ranges between 285 – 345.

Boxplot of Price Vs Garage

There are 4 boxplots generated representing distribution of number of Garage corresponding to overall house price.

- For **Garage 0**, the mean is closer to 1st quartile and the distribution is right skewed with 1 outlier. House Sale price ranges between 185 - 388 and there is one outlier.
- For **Garage 1**, mean is closer to the 1st quartile and is right skewed. House price ranges between 155.5 – 385.5.
- For **Garage 2**, Minimum is close to the first quartile, indicating that there are very less values in this region. Maximum is far from 3rd quartile hence the distribution is right skewed. House price ranges between 195 - 450 and highest house price among all garage size.
- For **Garage 3**, min value is at 1st quartile and maximum at 3rd quartile. Mean is at center. The distribution is right skewed. House price is between 299 & 339.9 and has lowest number of houses.

Boxplot of Price Vs School

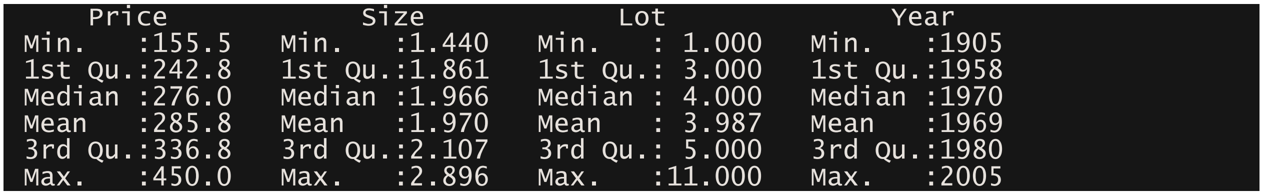
There are 6 boxplots generated representing distribution of School corresponding to overall house price.

- For **Alex**, the mean is closer to 1st quartile and the distribution is right skewed.House price ranges between 155.5 - 350 and has lowest price house.
- For **High**, the mean is closer to 1st quartile and the distribution is right skewed. House price ranges between 235 - 450 and has highest house price.
- For **NotreDame**, mean is closer to the 3rd quartile. Also, the maximum is very close to the 3rd quartile as compared to the distance of minimum from 1st quartile. House price ranges between 249.9 – 359.9.
- For **StLouis**, It shows almost symmetric shape hence symmetric normal distribution. House price ranges between 185 - 355 and has one house price outlier.
- For **StMarys**, House price ranges between 189.5 - 435 and has one house price outlier.
- For **Startford**, mean is closer to the 1st quartile. Also, the maximum is close to the 3rd quartile as compared to the distance of minimum from 1st quartile. House price ranges between 222.5 -349.5.

3: Using the summary, correlation and the pairs plots discuss the relationship between the response sales price and each of the numeric predictor variables.

- Summary of numeric variables Price, Size, Lot & Year:

```
House %>%

  select(Price,Size,Lot,Year) %>%

  summary()
```

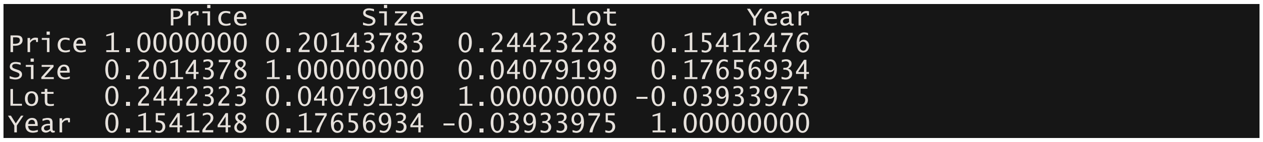


```
     Price            Size            Lot             Year     
 Min.   :155.5   Min.   :1.440   Min.   : 1.000   Min.   :1905  
 1st Qu.:242.8   1st Qu.:1.861   1st Qu.: 3.000   1st Qu.:1958  
 Median :276.0   Median :1.966   Median : 4.000   Median :1970  
 Mean   :285.8   Mean   :1.970   Mean   : 3.987   Mean   :1969  
 3rd Qu.:336.8   3rd Qu.:2.107   3rd Qu.: 5.000   3rd Qu.:1980  
 Max.   :450.0   Max.   :2.896   Max.   :11.000   Max.   :2005  
```

Summary provides min, 1st quantile, median, mean, 3rd quantile and max details of each of the numeric variables.

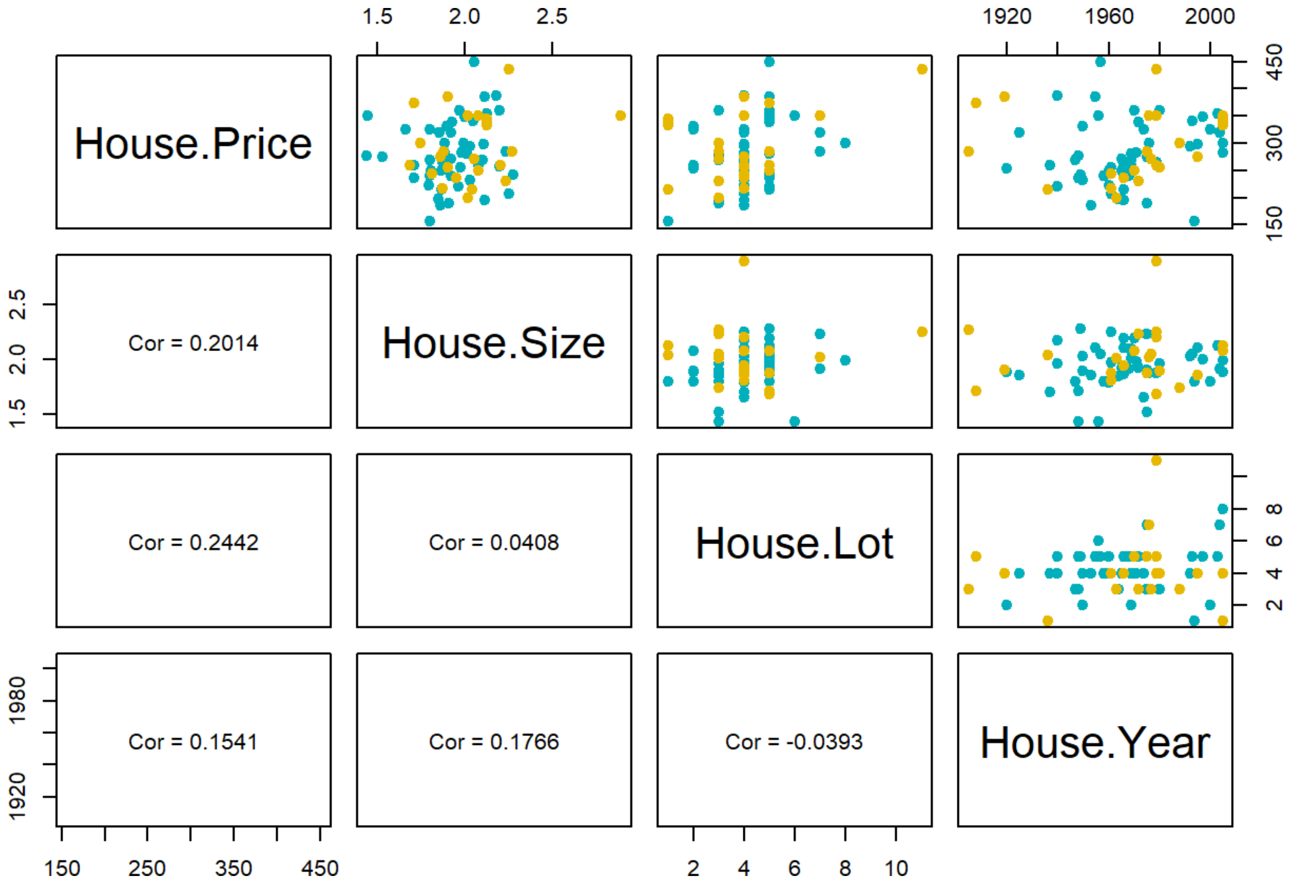
- Correlation among numeric variables Price, Size, Lot & Year:

```
House %>%

  select(Price,Size,Lot,Year) %>%

  cor()
```



```
          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
```

Below figure shows scatter plot of House Price with numerical variable: Price, Size, Lot and Year, along with their respective correlation:



The pairs function generates scatterplot of all continuous variables against the target variable Price. From the matrix of graphs, we can observe that, Price does not have a clear positive linear relationship with Size, Lot and Year.

Correlation matrix explains if the variables have positive or negative relationships. Also, it tells about the strength of the relationship whether its strong, moderate or weak.

Correlation with **Price** (target/dependent variable) :

- Positive weak relation with Size (20.14 %)
- Positive weak relation with Lot (24.42 %)
- Positive weak relation with weight (15.41 %)

Hence there is no numeric predictor variable with high correlation with house price.
Highest correlation is with the Lot variable of 0.244.

Other observed correlations :

- Size has Positive weak relation with Lot (4.08%)
- Lot has Negative weak relation with Year (-3.93%)
- Size has Positive weak relation with Year (17.66%)

Regression Model:

1. Fit a multiple linear regression model to the data with sales price as the response and size, lot, bath, bed, year, garage and school as the predictor variables. Write down the equation for this model.

Y Price = Beta0 + Beta1Size + Beta2Lot + Beta3Bath + Beta4Bed + Beta5Year + Beta6Garage + Beta7School + Error

```
lm(House$Price ~ (House$Size + House$Lot + House$Bath + House$Bed + House$Yea
r + House$Garage + House$School), data = House)
```

- Reference model is **Beta0 +Bath1 +Bed2 + Garage0 + School Alex** which is the base model for our regression model.
- In summary as well these predictor variables are hence not present and it tells us that the house price for a 1 bath, 2 bed, 0 garage and location near School Alex as the reference variables on which our regression model is built.

2. Interpret the estimate of the intercept term:

- Beta0 has Negative intercept value (-**884.3531)** is the expected value of House Price (dependent variable) when all independent/predictor variables (Size, Lot, Bath, Bed, Year, Garage and School) are set to 0.
- As we are getting negative intercept value for Beta0 as some of continuous independent variable have values far off from zero, we will rescale the continuous predictor variables by creating new predictor variables by subtracting each variable with its mean, and then fitting the regression model.
- Now we are getting intercept **376.10** which is positive and interpretable.

3. Interpret the estimate of size the parameter associated with floor size :

- Estimate of change in house price for unit increase in size is **59.4503** with p-value **0.045**(i.e. less than 0.05) with a condition of all other predictor variables kept constant and hence size variable will be significant, and we can reject null hypothesis for it.

4. Interpret the estimate of Bath1.1 the parameter associated with one and a half bathrooms.:

- Estimate of change in house price for unit increase in Bath1.1 is **135.8983** with p-value **0.007**** (i.e. less than 0.05) with a condition of all other predictor variable kept constant and thus we can say that it is highly related to the dependent variable Price.

5. Discuss and interpret the effect the predictor variable bed on the expected value of the house prices.

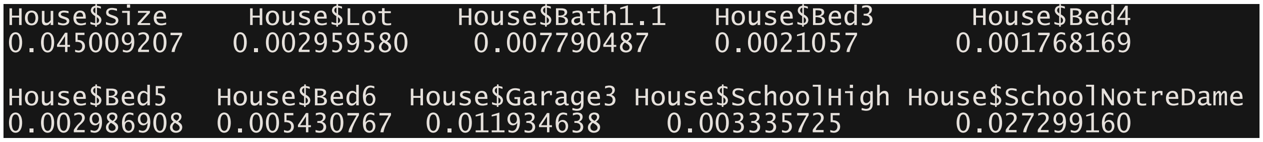
As mentioned earlier our reference model has Bed2 as base and hence with increase in below bed size the price of house will vary as follows:

- Bed3: For this variable house price will drop by 228.1052
- Bed4: For this variable house price will drop by 238.2609
- Bed5: For this variable house price will drop by 237.6155
- Bed6: For this variable house price will drop by 255.0211

The house price tends to decrease as the number of bedrooms are increased above 2.

6. List the predictor variables that are significantly contributing to the ex-pected value of the house prices –

Variables having their p value less than 0.05 are significantly contributing to the ex-pected value of the House Prices and those are as shown below:



```
House$Size      House$Lot    House$Bath1.1   House$Bed3      House$Bed4
0.045009207   0.002959580    0.007790487    0.0021057      0.001768169

House$Bed5   House$Bed6  House$Garage3 House$SchoolHigh House$SchoolNotreDame
0.002986908  0.005430767  0.011934638    0.003335725       0.027299160
```

7. For each predictor variable what is the value that will lead to the largest expected value of the house prices.

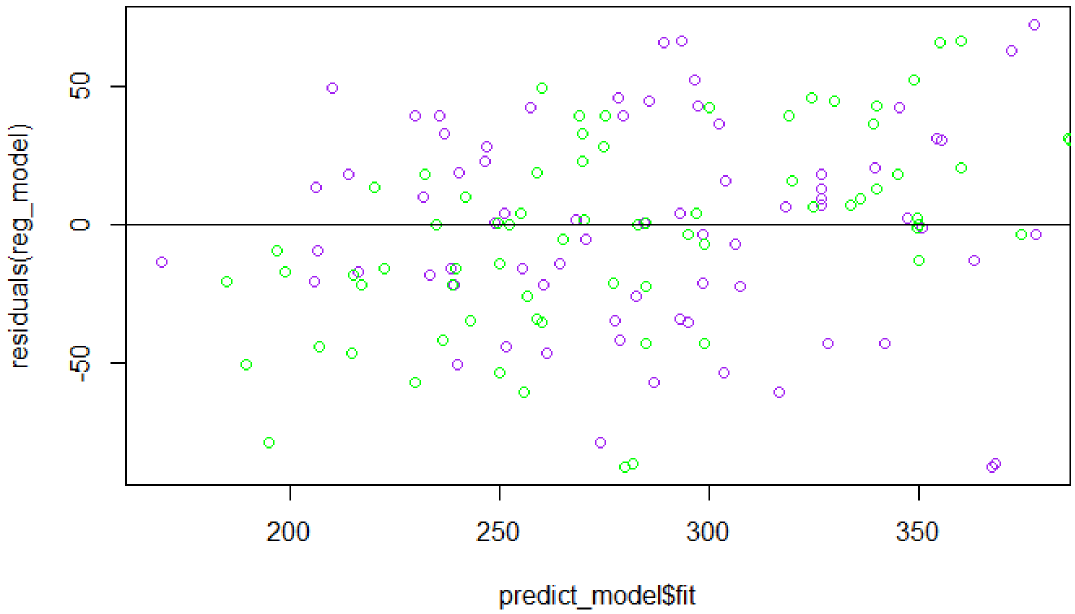
Size = 2.896, Lot = 11, Bath = 1.1, Bed = 2, Garage = 2, School = High, Year = 2005 will contribute to the highest expected value of the house price.

8. For each predictor variable what is the value that will lead to the lowest expected value of the house prices.

Size = 1.44, Lot = 1, Bath = 1, Bed = 6, Garage = 3, School = Alex, Year = 1905 will contribute to the lowest expected value of the house price.

9. By looking at the information about the residuals in the summary and by plotting the residuals do you think this is a good model of the expected value of the house prices.

The Residual Standard Error is the average amount that the response will deviate from the true regression line. When the residual standard error is exactly 0, the model fits the data perfectly. Residual standard error is 42.13 on 55 degrees of freedom which means it’s not a perfect fit, but data points are spread across the line and we can say that it is an almost good model.



10. Interpret the Adjusted R-squared value.:

The adjusted R-squared value is 0.5125 and is decreasing (with respect to R square value of 0.6425) . The higher the adjusted R-squared value the better is the model.
House Price is explained 51.25% by the included predictor variables.

11. Interpret the F-statistic in the output in the summary of the regression model.

F test: Indicates whether regression model provided better fit to the data than a model that contains no independent variables.

Hypothesis tested is Null Hypothesis(H0) and Alternate Hypothesis(Ha)

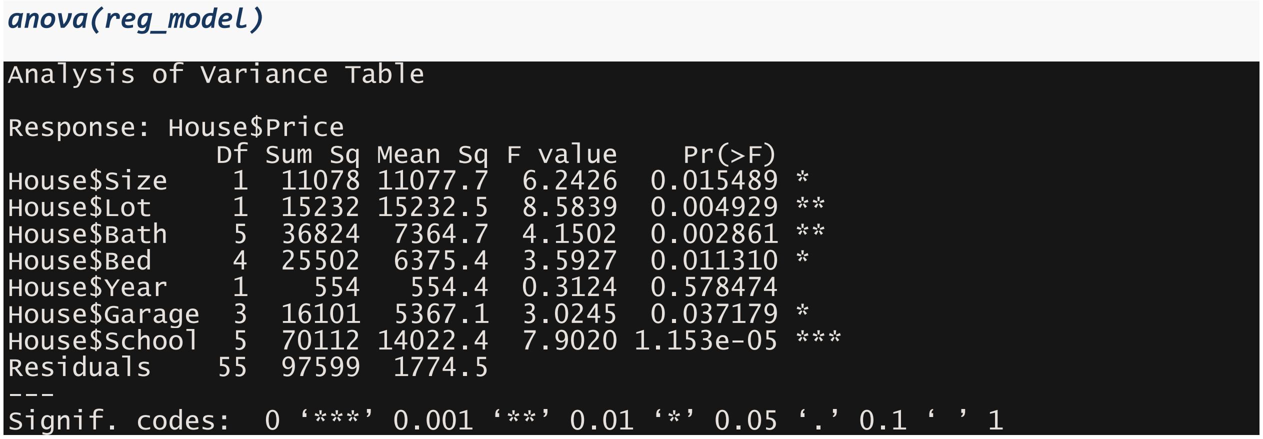
- **Null Hypothesis(H0):** The model with no independent/predictor variables fits the data and adding them doesn't improve the model(all predictor variables are zero)
- **Alternate Hypothesis(Ha):** Model fits the data better than the intercept-only model by adding the predictor variables(at least one of the predictors is non zero)

F-statistic value is 4.942 on 20 and 55 DF, p-value is 1.265e-06(less than 0.05) and thus we can reject the Null Hypothesis(H0) and model improves with predictor variable addition.

ANOVA:

1. Compute the type 1 anova table. Interpret the output. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem.

Analysis of Variance (ANOVA) consists of calculations that provide information about levels of variability within a regression model and form a basis for tests of significance.



```
anova(reg_model)
```

```
Analysis of Variance Table

Response: House$Price
             Df Sum Sq Mean Sq F value    Pr(>F)    
House$Size    1  11078 11077.7  6.2426  0.015489 *  
House$Lot     1  15232 15232.5  8.5839  0.004929 ** 
House$Bath    5  36824  7364.7  4.1502  0.002861 ** 
House$Bed     4  25502  6375.4  3.5927  0.011310 *  
House$Year    1    554   554.4  0.3124  0.578474    
House$Garage  3  16101  5367.1  3.0245  0.037179 *  
House$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
```

Hypothesis being tested is NULL and Alternate Hypothesis:
a) **Null Hypothesis (H0) :** No relation between dependent Price and Independent variables(all of them are zero.). Means between groups are same.
b) **Alternate Hypothesis (Ha) :** Independent variable did affect the dependent variable Price and contribute to the model(all predictor variable are not zero). At least mean of one group is different.

From above **Type I anova** table we can interpret that:

- School is the most significant contributor in the regression model as its P value
- (1.153e-05 ***) is way less than 0.05 and indicates strong evidence against the null hypothesis.
- Lot(0.004929 **) & Bath(0.002861**) are also significantly contributing towards the model.
- Size, Bed & Garage are slightly contributing to the model as their p value are close to 0.05

2. Which predictor variable does the type 1 anova table suggest you should remove the regression analysis.

The P-value for **Year** is **0.578474** , which is more than 0.05 & F statistic of low value 0.031 and thus we fail to reject null hypothesis for it and can say that there is a weak relationship between predictor variable year and house price. Hence, we should remove Year to make a better model.

3. Compute a type 2 anova table comparing the full model with all predictor variables to the reduced model with the suggested predictor variable identi_ed in the previous question removed.

We have created Reduced model by removing Year variable on the basis of previous conclusion.

Now using this **reg_reduced_model** with our original **reg_model** to perform Anova type II test.

```
reg_reduced_model= lm(House$Price~House$Size+House$Lot+House$Bath+House$Bed+H
ouse$Garage+House$Sc hool,data=House)

anova(reg_model, reg_reduced_model)
```

```
Analysis of Variance Table

Model 1: House$Price ~ (House$Size + House$Lot + House$Bath + House$Bed +
    House$Year + House$Garage + House$School)

Model 2: House$Price ~ House$Size + House$Lot + House$Bath + House$Bed +
    House$Garage + House$School

   Res.Df  RSS   Df  Sum of Sq    F     Pr(>F)
1     55  97599
2     56 102402 -1    -4802.6    2.7064 0.1057
```

Hypothesis being tested is NULL and Alternate Hypothesis:

a) **Null Hypothesis (H0) :** No relation between dependent Price and Independent variable Year(BetaYear=0) and is not significantly contributing to the model.
b) **Alternate Hypothesis (Ha) :** Independent variable Year did affect the dependent variable Price and contribute to the model(BetaYear is not equal to 0).

From above **Type II anova** table we can interpret that:
P value is 0.1057 which is much higher than 0.05 and hence we fail to reject null hypothesis for the variable **year** .

Diagnostics:

1. Check the linearity assumption by interpreting the added variable plots and component-plus-residual plots. What effect would non-linearity have on the regression model and how might you correct or improve the model in the presence of non-linearity?

Added variable plots attempt to show the effect of adding an additional variable to the model (given that one or more independent variables are already in the model).

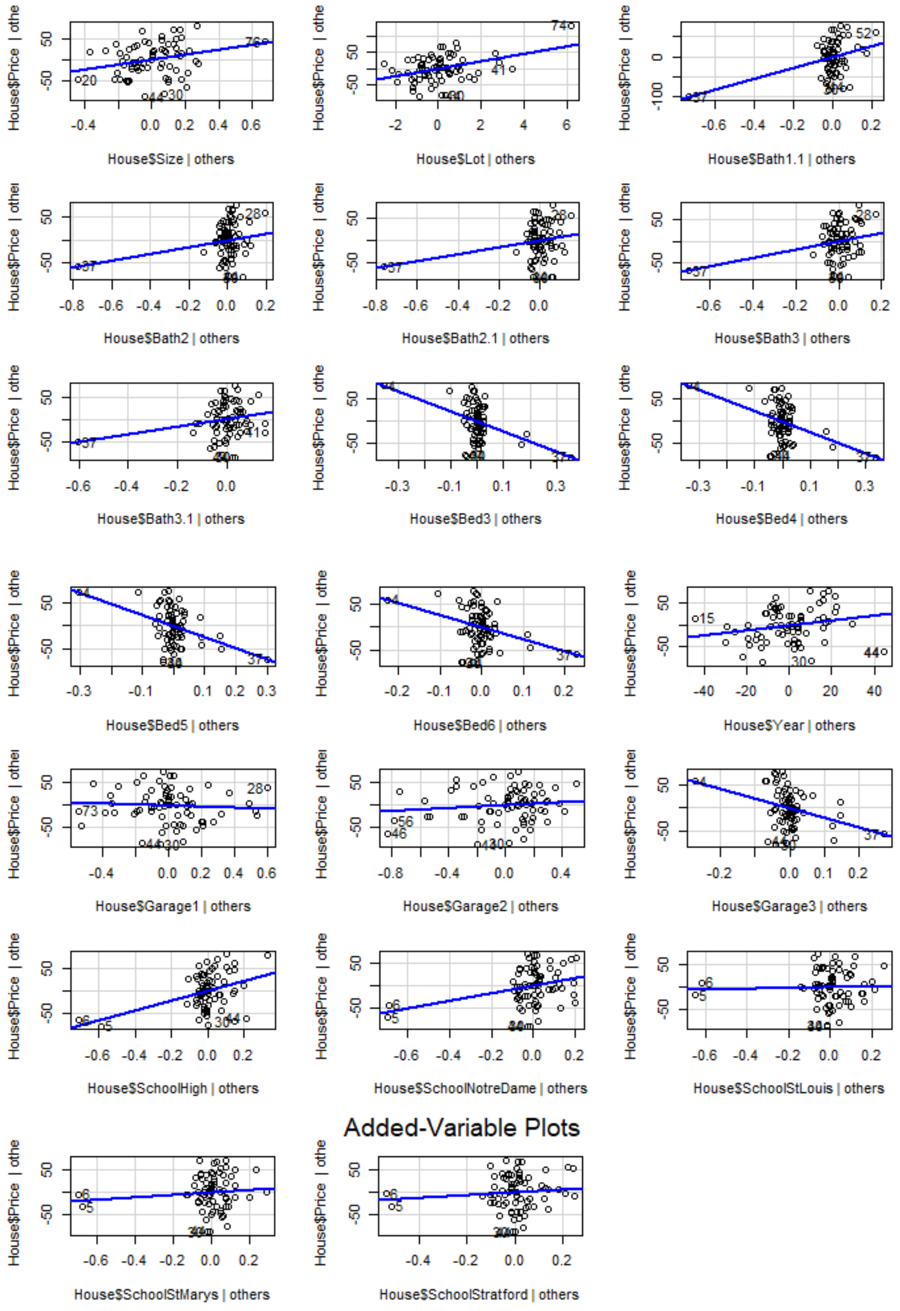
a) Adding **Size** to the model given that rest all variables are already in the model does have a significant impact as the slope of the fitted line is different from zero. There are possible outliers at 20,44,30. As the size of the house increases the variability in the price also increases
b) Adding **Lot** to the model given that rest all variables are already in the model. There are possible outliers at 25,30,41
c) Adding **Bath(1.1,2,2.1,3,3.1)** to the model given that rest all variables are already in the model. There are possible outliers at 25 and 30
d) Adding **Bed(3,4,5,6)** to the model given that rest all variables are already in the model. There are possible outliers at 25.
e) Adding **Garage(1,2,3)** to the model given that rest all variables are already in the model. There are possible outliers at 25 and 30.
f) Adding **School** to the model given that rest all variables are already in the model. There are possible outliers at 25 and 30.

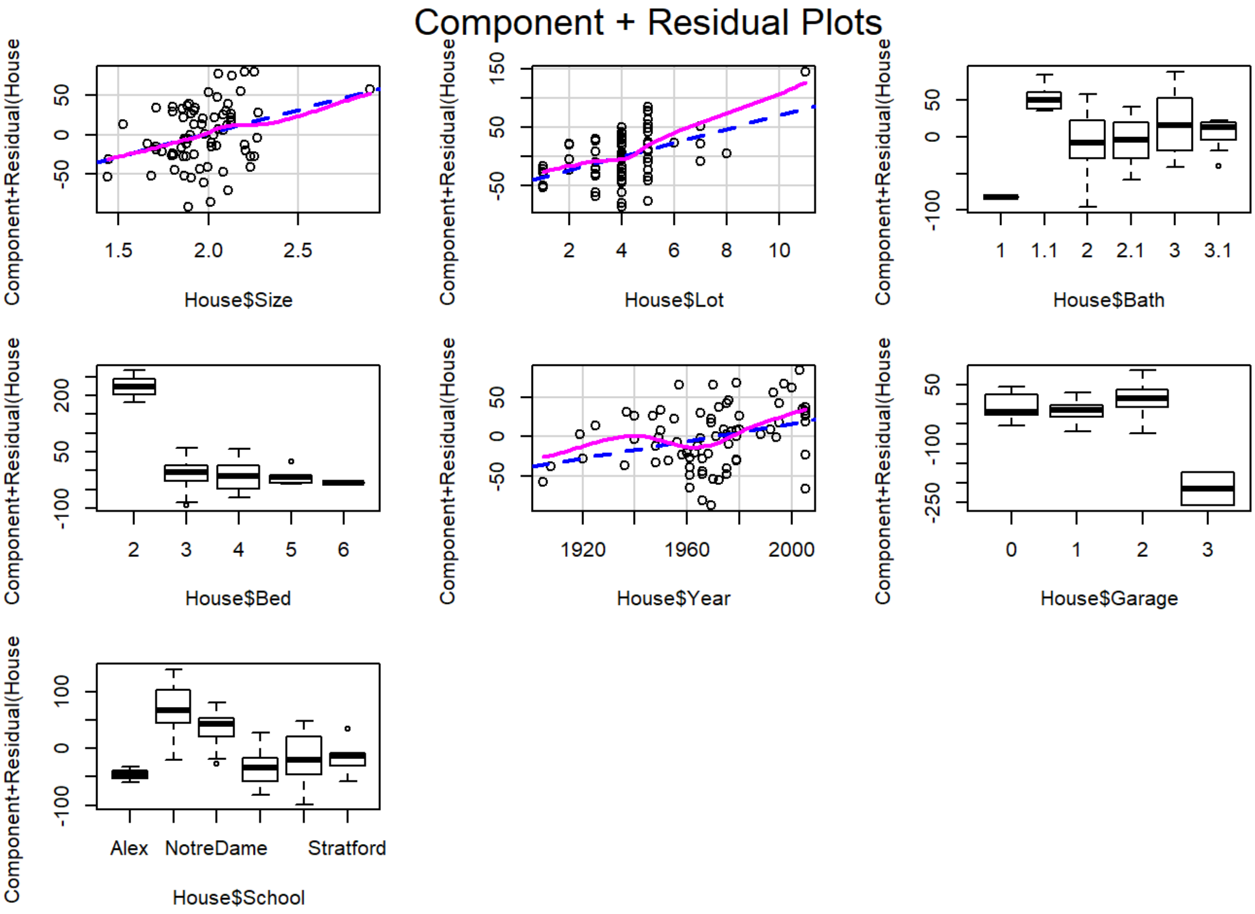
```
avPlots(reg_model)
```

Component Residual Plot:

a) The relationship between House Price and Size is approximately linear as the dashed and the pink line do not differ dramatically.
b) The relationship between House Price and Lot is approximately linear as the dashed and the pink line do not differ dramatically.
c) The relationship between House Price and Bath seems to be strongly linear.
d) The relationship between House Price and Bed is approximately linear as the dashed and the pink line differ.
e) The relationship between House Price and Year is approximately linear as the dashed and the pink line differ but not dramatically.
f) The relationship between House Price and Garage is approximately linear as the dashed and the pink line do not differ dramatically.
g) The relationship between House Price and School seems to be approximately linear.

```
crPlots(reg_model)
```


Added-Variable Plots
House$Price | othe
House$Size | others
House$Lot | others
House$Bath1.1 | others
House$Bath2 | others
House$Bath2.1 | others
House$Bath3 | others
House$Bath3.1 | others
House$Bed3 | others
House$Bed4 | others
House$Bed5 | others
House$Bed6 | others
House$Year | others
House$Garage1 | others
House$Garage2 | others
House$Garage3 | others
House$SchoolHigh | others
House$SchoolNotreDame | others
House$SchoolStLouis | others
House$SchoolStMarys | others
House$SchoolStratford | others



Conclusion:

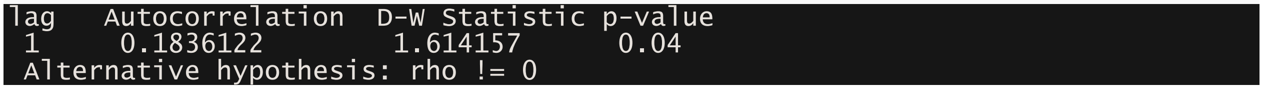
- Variable Lot, Size, Bath, School High and Notre Dame School variables show strong positive linear relationship.
- Garage3 and Bed 3,4,5,6 exhibit strong negative linear relationship.
- Year, Garage 1, Garage 2 and School St Louis , St Marys , Startford doesn't show strong linear relationships.

If we have non-linearity, then the model that will be built will be biased and will have the inconsistent Beta estimates and the outcome does not change in proportion to a change in any of the inputs.

- We can use added variable plots, scatterplots, component+ residual plots to identify the non-linear relationship between response and predictor variables.
- Correction and improving the model by transformations, splines, polynomials.
- We can use different types of regression which cover non-linear relationship between variables (based on e.g. exponential growth models, logistic model, exponential decay model etc.).

2. Check the random/i.i.d. sample assumption by carefully reading the data description and computing the Durbin Watson test (state the hypothesis of the test, the test statistic and p-value and the conclusion in the context of the problem). What are the two common violations of the random/i.i.d. sample assumption? What e_ect would dependant samples have on the regression model and how might you correct or improve the model in the presence of dependant samples?

```
dwt(reg_model)
```



```
lag   Autocorrelation  D-W Statistic p-value
 1     0.1836122        1.614157      0.04
 Alternative hypothesis: rho != 0
```

- As per **DWT** the test statistic lies between 0 and 4. If the Durbin–Watson statistic is substantially less than 2, there is evidence of positive serial correlation. If Durbin–Watson is less than 1.0, there may be cause for alarm and p-value should be above 0.05

- The Null and Alternate hypothesis of the DWT:

1. H0 - There is no autocorrelation in the regression model.
2. Ha - There is autocorrelation in the regression model.

- The Durbin Watson test statistic is 1.614 and the p-value is 0.04 so the hypothesis of no autocorrelation is rejected, and the observations cannot be classed as independent.

- The common violations in it are the recurring measurements present on the different timelines, several measurements of the identical subject and if the observations can be broken down in meaningful groups. This would result in the non-constant variance; structure dependence and outliers can cause inefficiency.
- To correct this, we can use mixed effect model

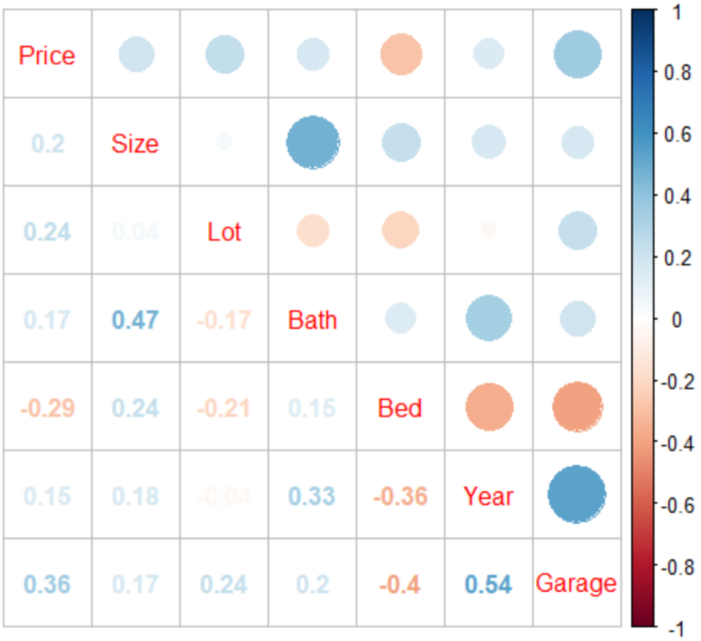
3. Check the collinearity assumption by interpreting the correlation and variance inflation factors. What effect would multicollinearity have on the regression model and how might you correct or improve the model in the presence of multicollinearity.

- There is multi-collinearity if there is a strong relationship between predictor variables i.e they are highly related to each other.

1. Correlation = 1, It means if one variable goes up by 1 then other will also go up by 1.
2. Correlation = 0, It means there is no correlation between two predictor variables.

- From below correlation plot we can say that there is a positive **correlation** between respective two variables, but it is weak and likely insignificant. We do not consider **correlations** significant until the value surpasses at least 0.8.

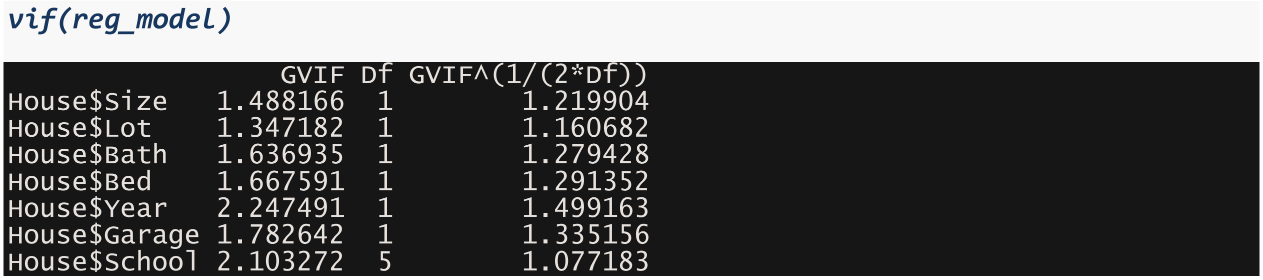
Correlation Plot:



- The **VIF** (variance inflation factor) in last column are approximately one indicating that we don't have a multicollinearity problem with a regression including all predictor variables .

1. **VIF = 1** , It means no correlation among the jth predictor and remaining predictor variables.
2. **VIF > 4** , It means warrant further investigations.
3. **VIF > 10** , It means serious multi collinearity is present and require correction.

- We can use Partial least square regression, Ridge Regression ,Principal Components analysis or remove highly correlated predictors to remove multi-collinearity.

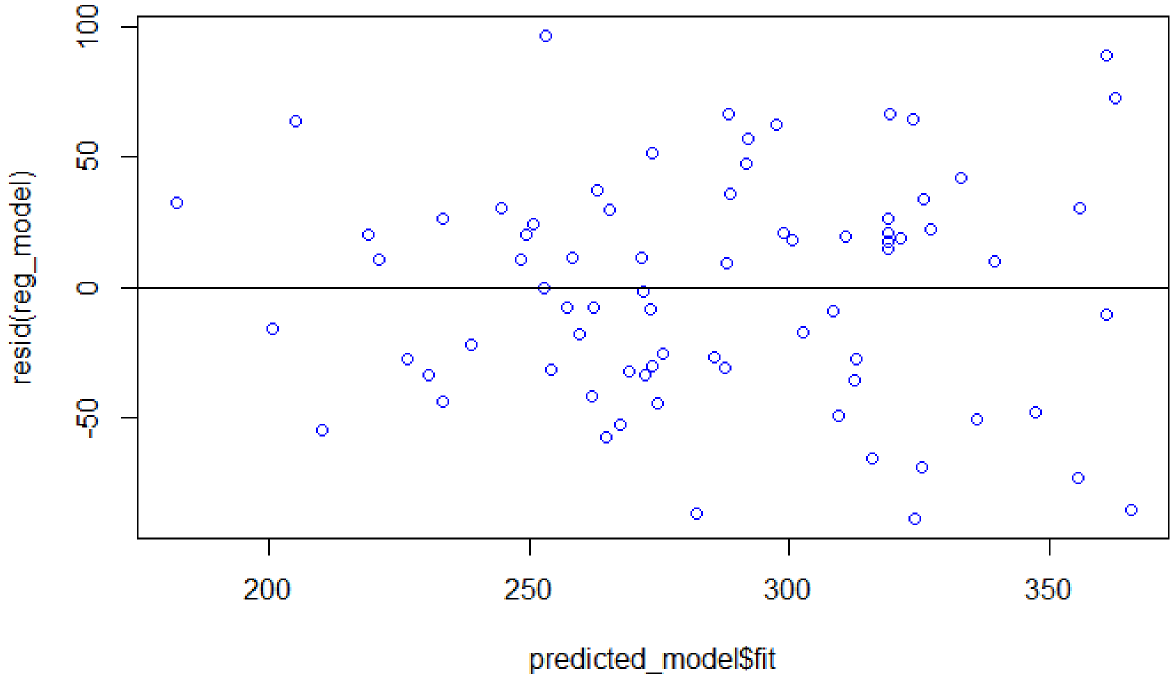


```
vif(reg_model)
```

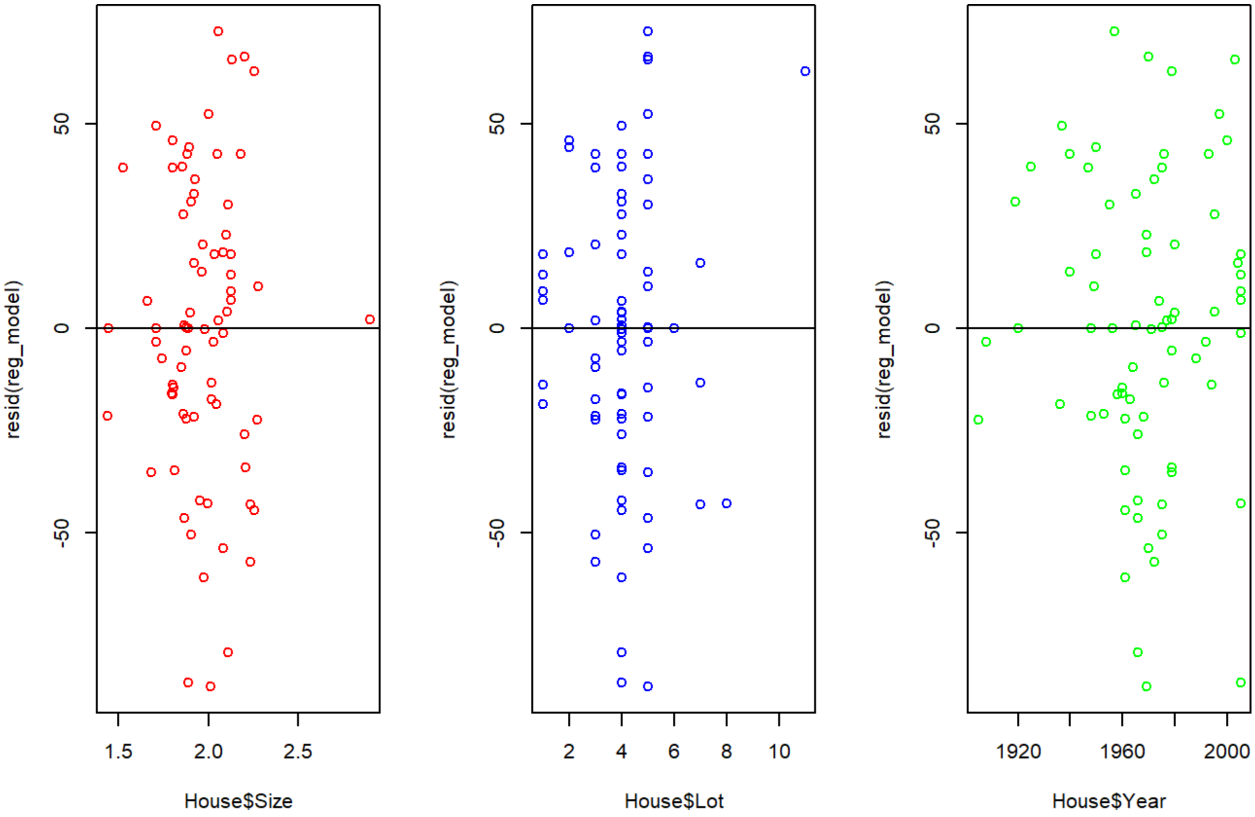
```
                  GVIF Df GVIF^(1/(2*Df))
House$Size    1.488166  1        1.219904
House$Lot     1.347182  1        1.160682
House$Bath    1.636935  1        1.279428
House$Bed     1.667591  1        1.291352
House$Year    2.247491  1        1.499163
House$Garage  1.782642  1        1.335156
House$School  2.103272  5        1.077183
```

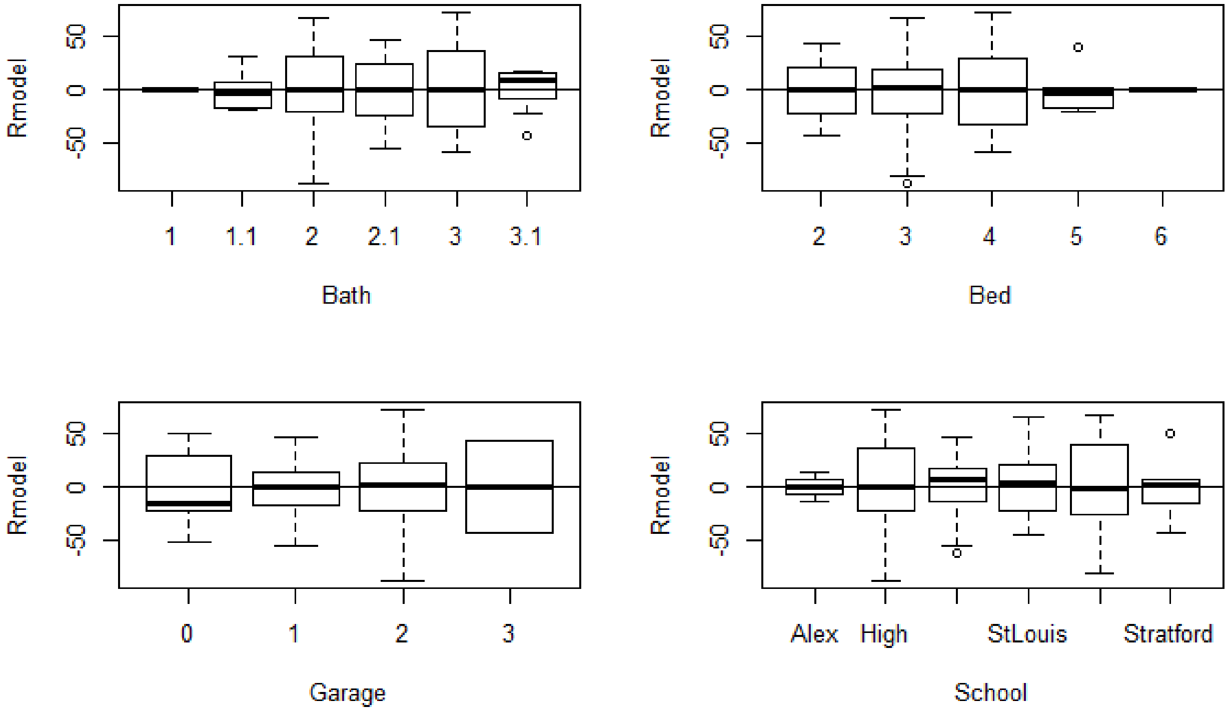
4. Check the zero conditional mean and homoscedasticity assumption by interpreting the studentized residuals vrs fitted values plots and the stu-dentized residuals vrs predictor variable plots. What effect would het-eroscedasticity have on the regression model and how might you corrector improve the model in the presence of heteroscedasticity.

Plots: From below residual vs predicted value plot we can interpret that there is constant variance and there is no funnel like shape present.



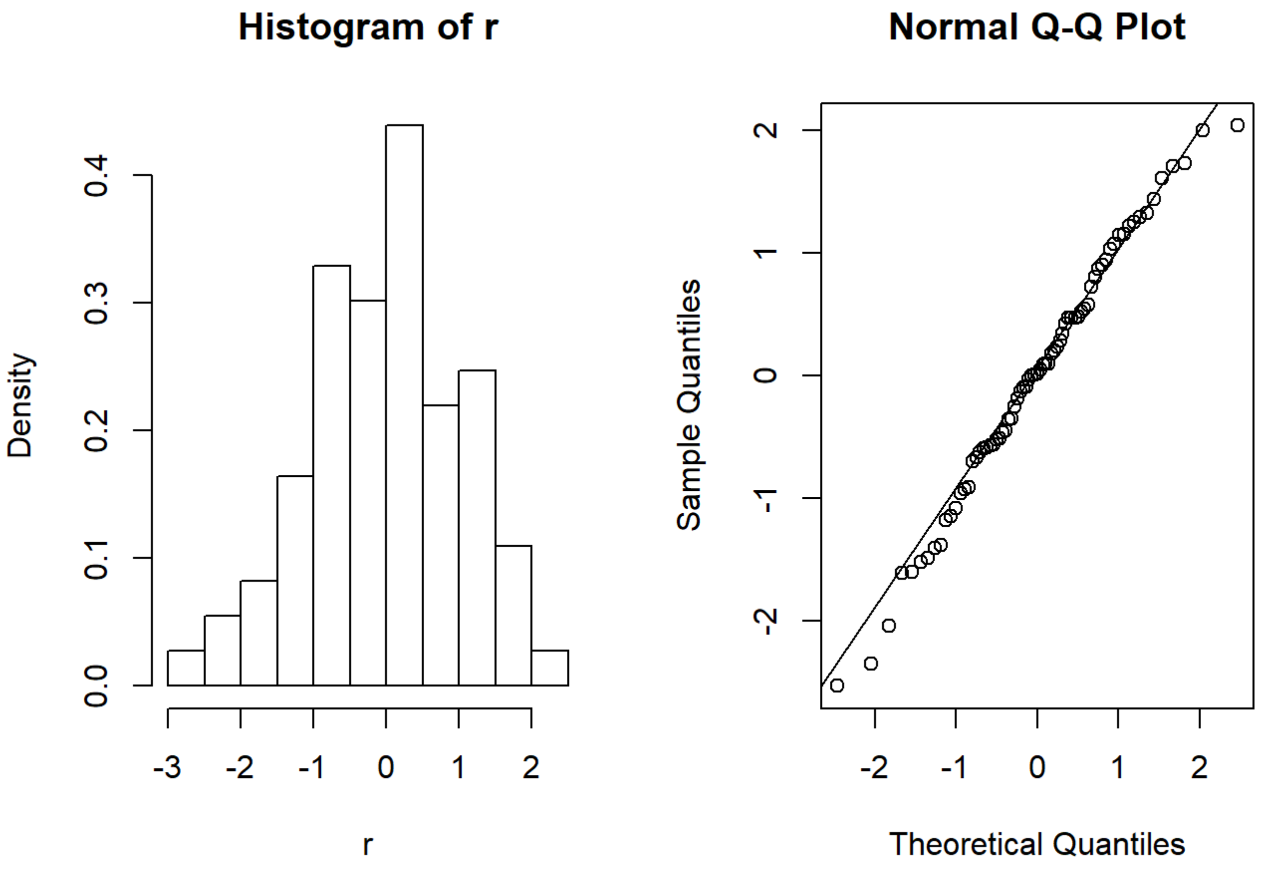
- In plots, there is symmetric pattern and across the mean zero have constant spread throughout the range.
- The effect is that the standard errors are biased or distorted.
- To Improve the model, we can use Weighted Least Squares method.


resid(reg_model)
50
0
-50
1.5
2.0
2.5
House$Size
resid(reg_model)
50
0
-50
2
4
6
8
10
House$Lot
resid(reg_model)
50
0
-50
1920
1960
2000
House$Year


Rmodel
50
0
-50
1
1.1
2
2.1
3
3.1
Bath
Rmodel
50
0
-50
2
3
4
5
6
Bed
Rmodel
50
0
-50
0
1
2
3
Garage
Rmodel
50
0
-50
Alex
High
StLouis
Stratford
School

5. Check the Normality assumption by interpreting the histogram and quantile-quantile plot of the studentized residuals. What effect would non-normality have on the regression model and how might you correct or improve the model in the presence of non-normality.

Histogram and QQ Plot:



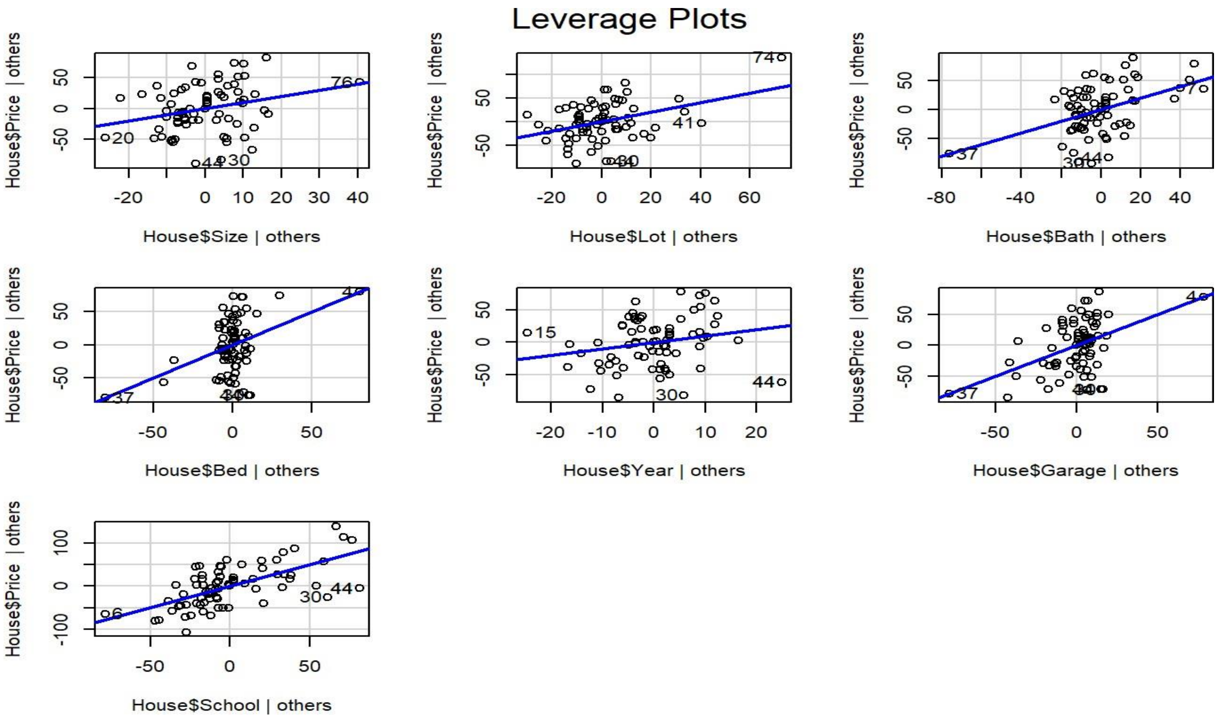
- Difference between what happened and what our model predicted is the Normality assumption.
- The histogram of studentized residuals have bell-shaped curve which means that residuals are normally distributed.
- QQ plot is useful for comparing the distributions. It Plots the quantile.
- In above QQ Plot the shown line tracks the residual points with some small deviation at the end which is expected and means that the residuals are normally distributed.
- If there is non-normality in the model, then Sampling distribution of Beta estimate will not be normal.
- Test statistics will not have t or F distributions correct.
- Probability of Type-I error will not be alpha. 1 alpha CI will not have 1-alpha coverage.
- To Improve the model in presence of non-linearity we can do transformations(response or predictor), residuals or use different models.

Leverage, Influence and Outliers:

1. What is a leverage point? What effect would a leverage point have on the regression model? Use the leverage values and the leverage plots to see if there is any leverage points.

- The leverage measures the amount by which the predicted value would change if the observation was shifted by one unit in the direction of y.
- Leverage always takes values between the range of 0 and 1.
- A point with zero leverage has no effect on the regression model. If a point has leverage equal to 1 the line must follow the point perfectly.
- If Y-Value follows the general trend of rest of the data, then leverage point has little effects on the estimates of regression coefficients and only affects the model summary statistics.
- If the Y-Value vary in the trend from the rest of the data, then leverage point affects the model summary statistics and estimates of regression coefficients.

```
leveragePlots(reg_model)
```

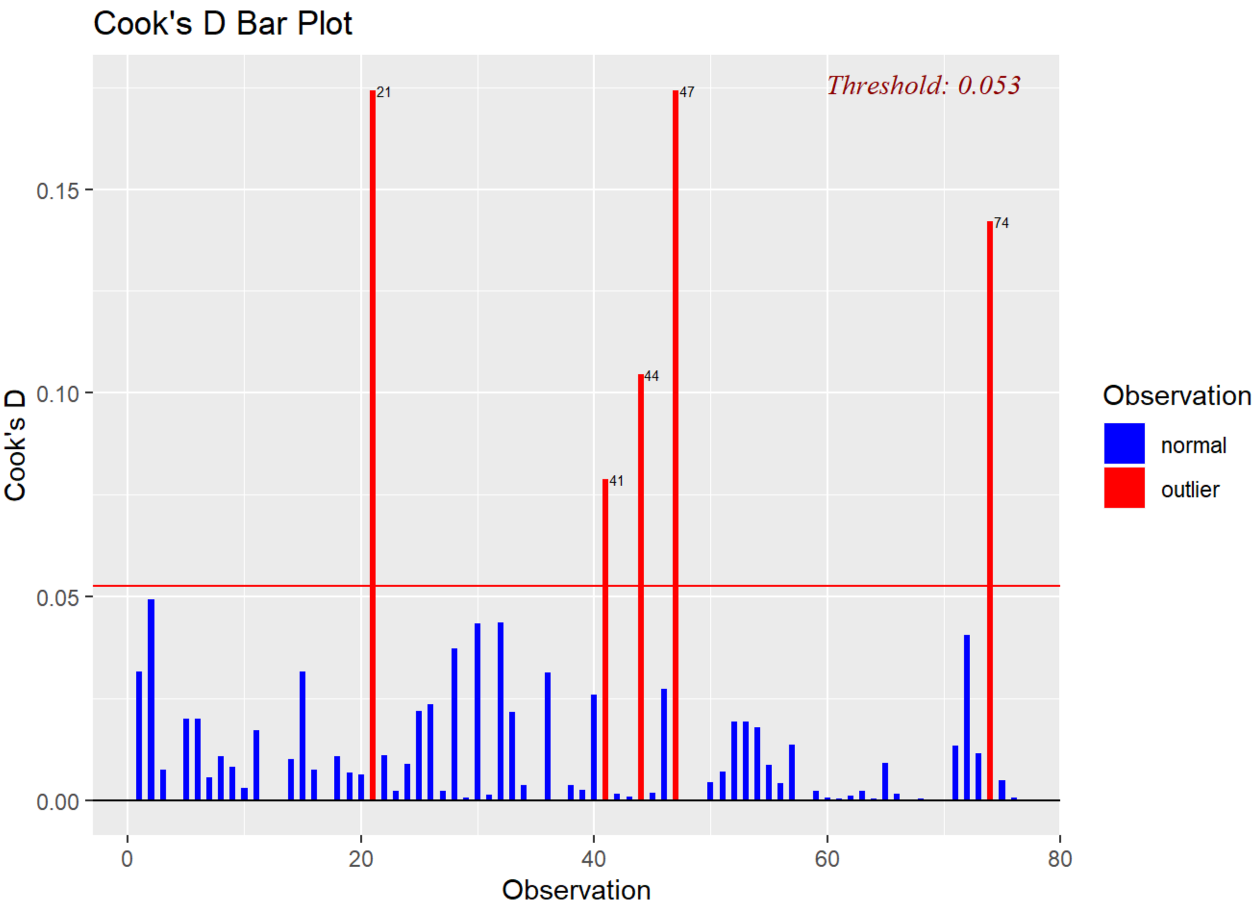


- A leverage point is a point that is distant from the bulk of the points can have a large influence on the parameter estimates for the term.
- From the above plot we can say that 6,20,76,41,74,37,4,4,15,30,44 seem little usual and can be said as leverage points.

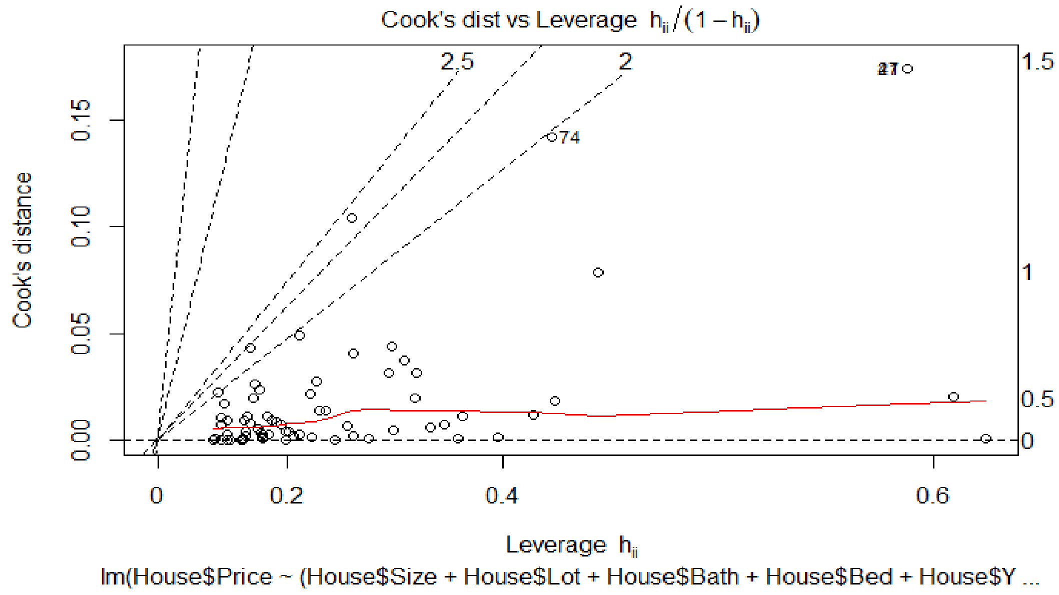
2. What is an influential point? What effect would an influential point have on the regression model? Use the influence plot to see if there is any influence points.

- An influential point is the point that is if removed from the data then it would significantly change the fit of the data.
- 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.
- We are here analyzing influential point using Cook's Distance. It suggests that if Cook's distance is greater than 1 then that point is influential and further investigation is required on it.

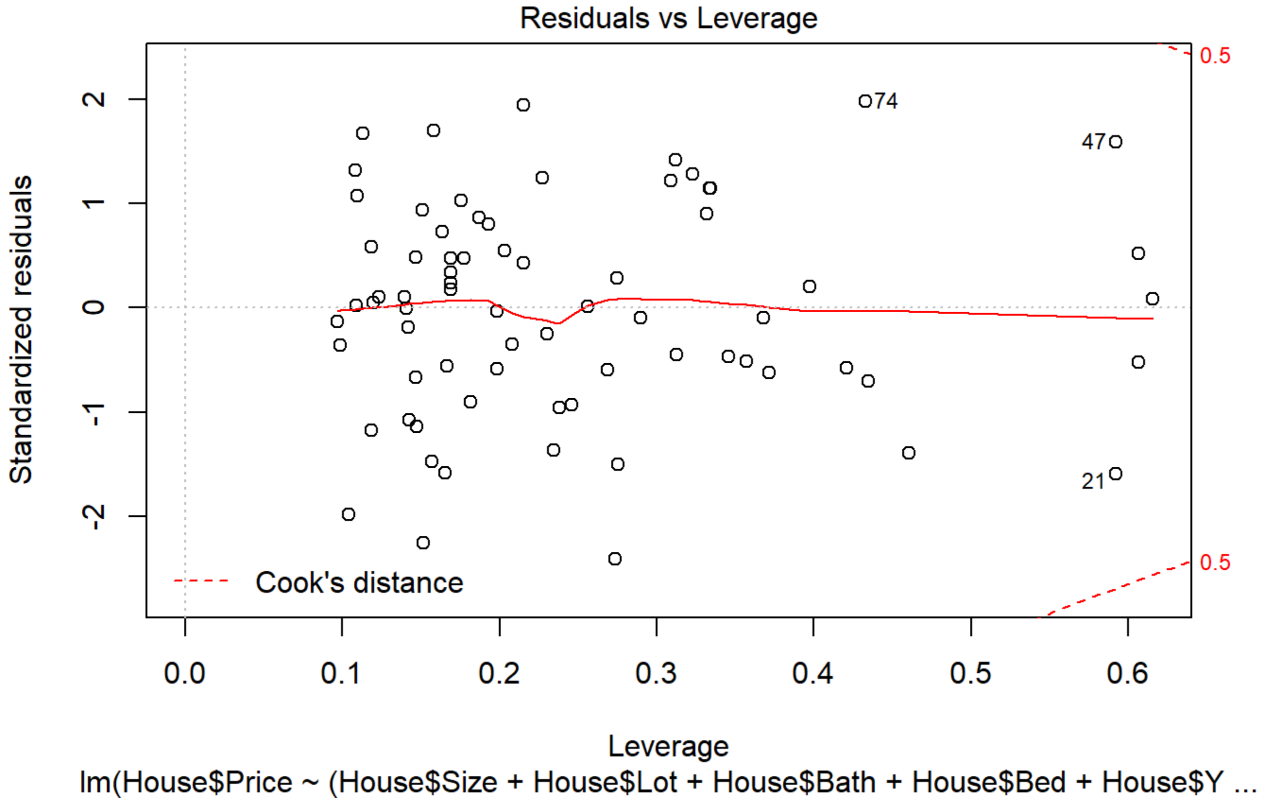
```
ols_plot_cooksd_bar(reg_model)
```



- Above plot shows that there is potential outlier in observations highlighted by the red lines and all the points are within Cook's distance.

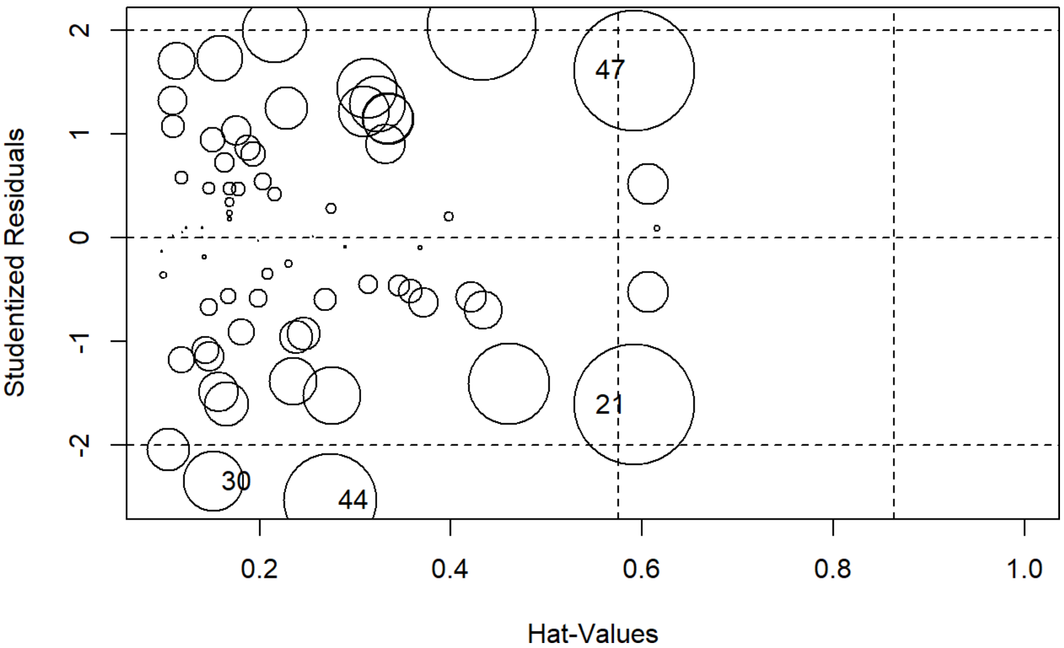


```
plot(reg_model,5)
```



- Above plots suggests that all the points within the Cook's distance .

```
influencePlot(reg_model)
```



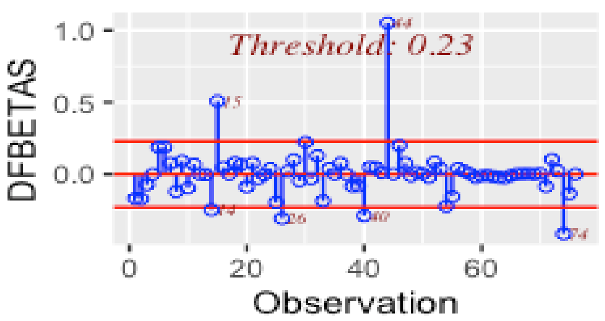
- The Cook's distance statistic is a measure, for each observation in turn, of the extent of change in Y hat when that particular observation is omitted. Any observation for which the Cook's distance is close to 1 or more, or that is substantially larger than other Cook's distances (highly influential data points), requires investigation.

- Above plot shows that the larger the size of circle, larger is the Cook's distance.

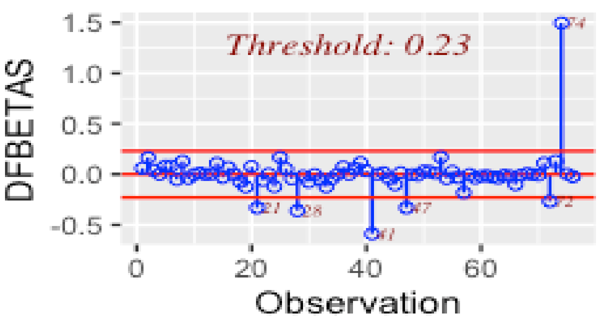
```
ols_plot_dfbetas(reg_model)
```

page 1 of 6

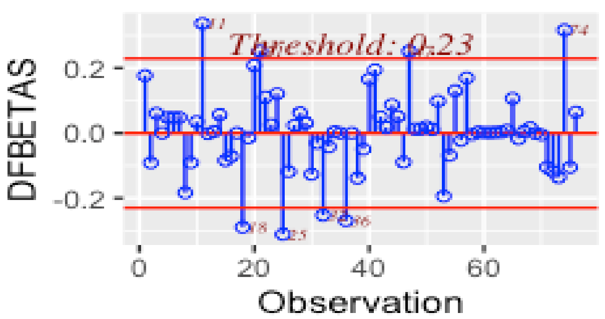
Influence Diagnostics for



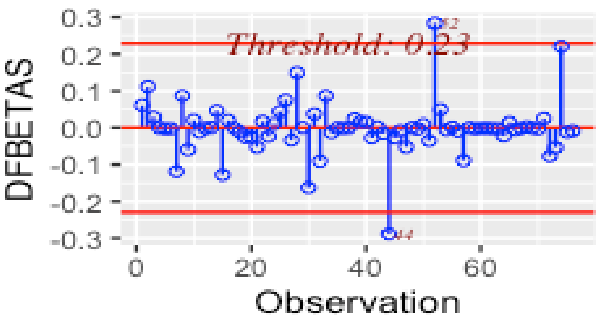
Influence Diagnostics fo



Influence Diagnostics fo

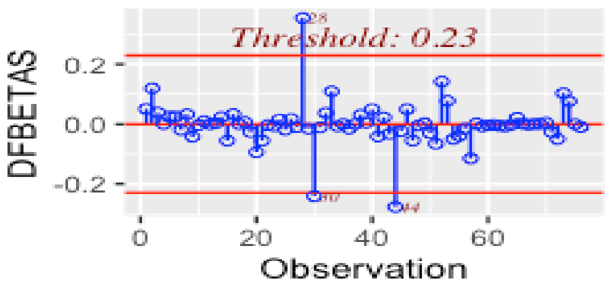


Influence Diagnostics fo

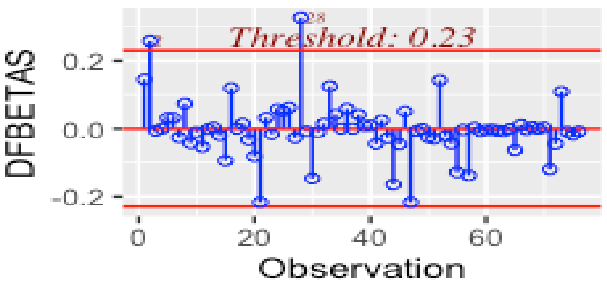


page 2 of 6

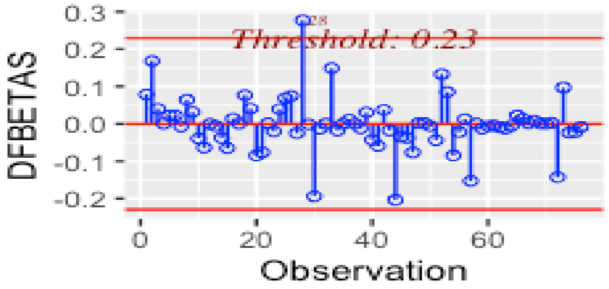
Influence Diagnostics fo



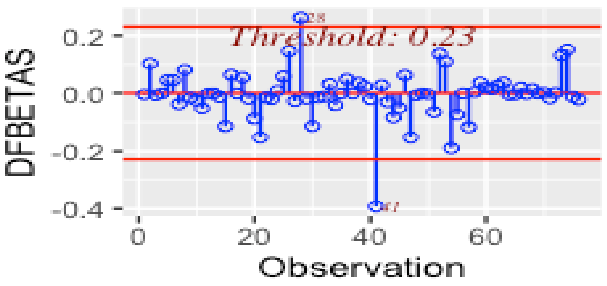
Influence Diagnostics fo



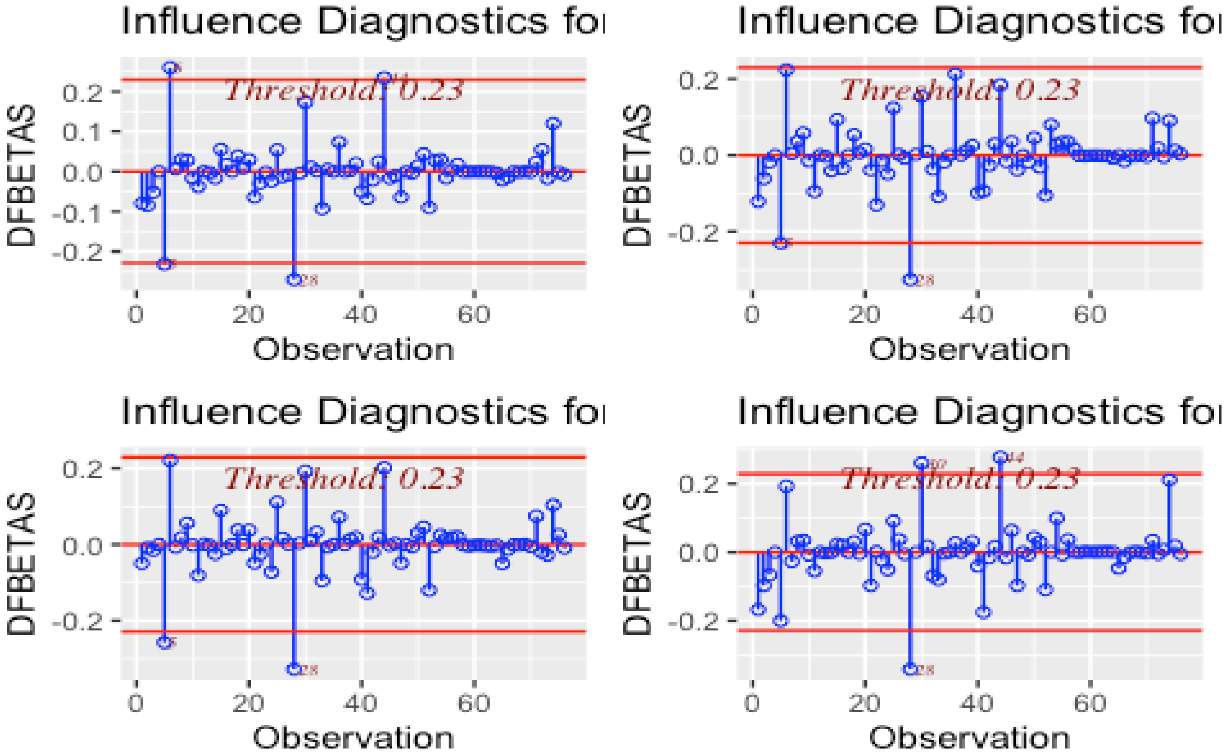
Influence Diagnostics fo



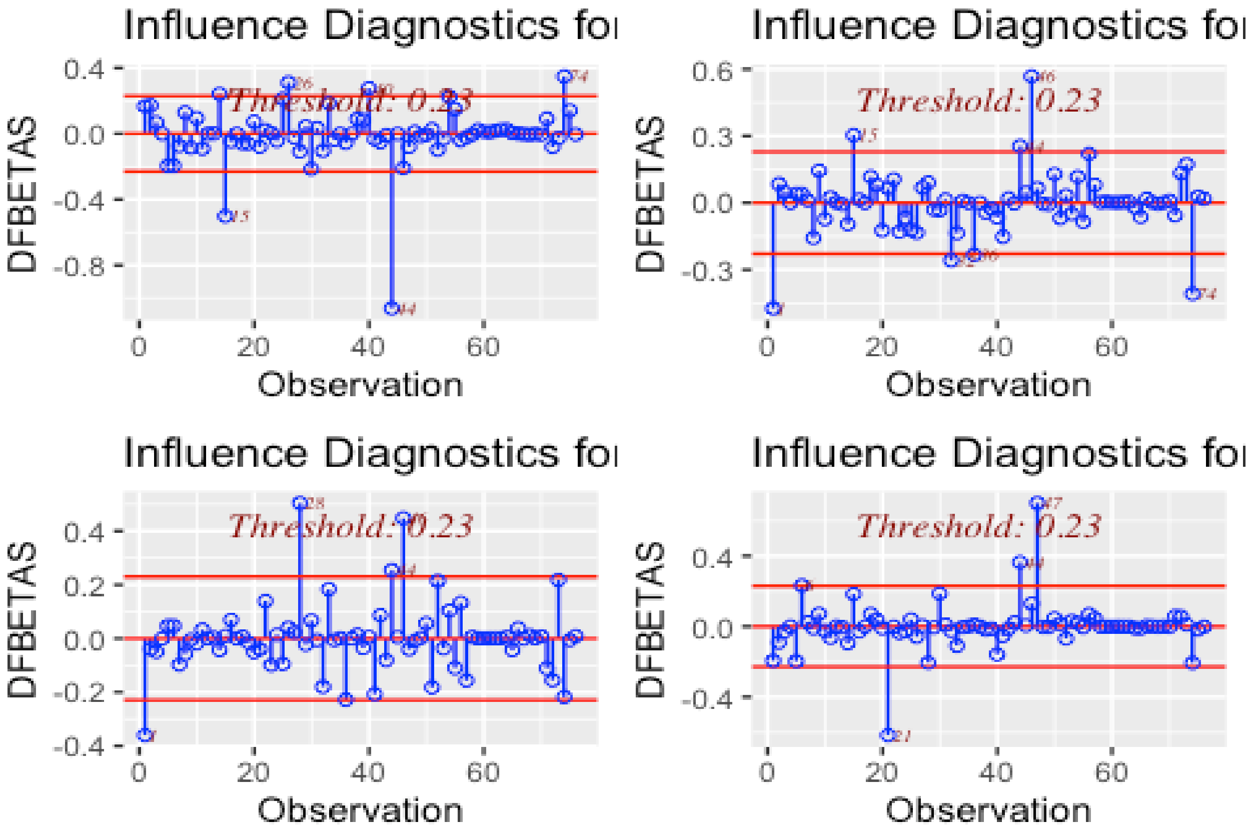
Influence Diagnostics fo

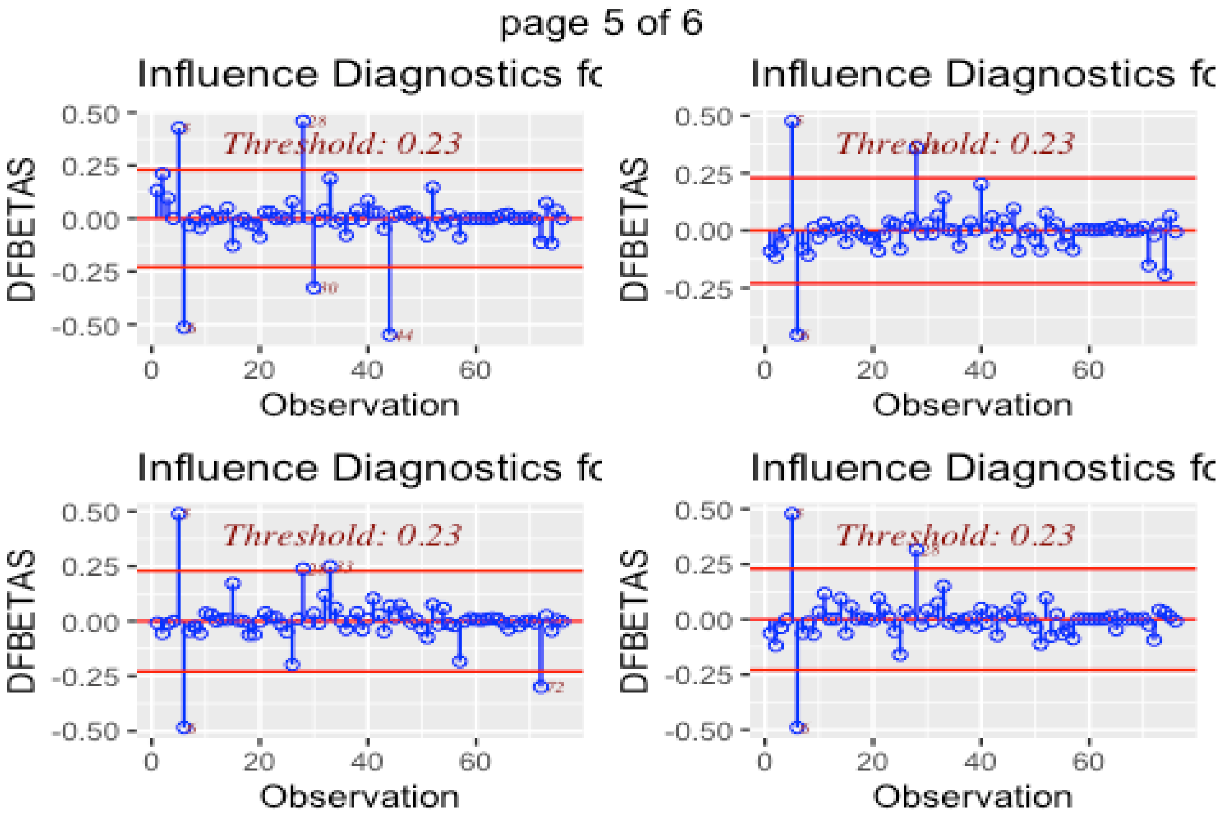


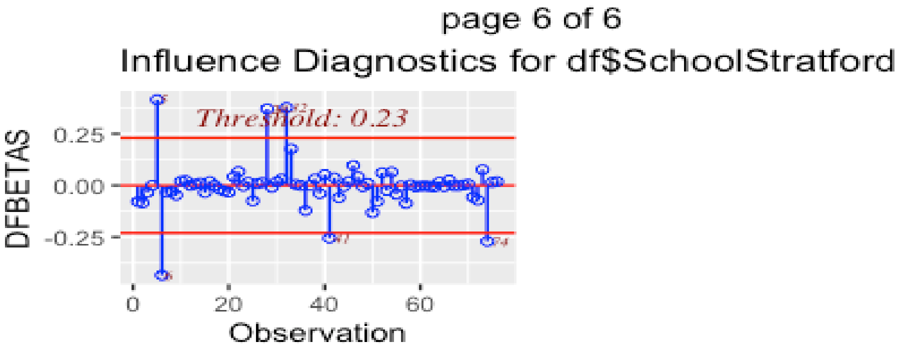
page 3 of 6



page 4 of 6







From the set of plots, we can say that it is for each coefficient "Bj" for each of the observations. It highlights the possible outliers in the graph that are potentially above the threshold value of 0.23.

3. What is an outlier? What effect would an outlier have on the regression model? How would you correct for outliers? Use the outlier test and outlier and leverage diagnostics plot to see if there are any outliers. Deal with the outliers if any are identified.

- An outlier is an observation where the response does not correspond to the fitted value of the model to the bulk of the data.
- An outlier can affect the model drastically and can cause a change to the model equation and in turn may result in bad prediction or estimation.
- An outlier can be detected by estimating the model with the help of other data apart from the outlier.

- It's important to investigate the nature of the outlier before deciding.

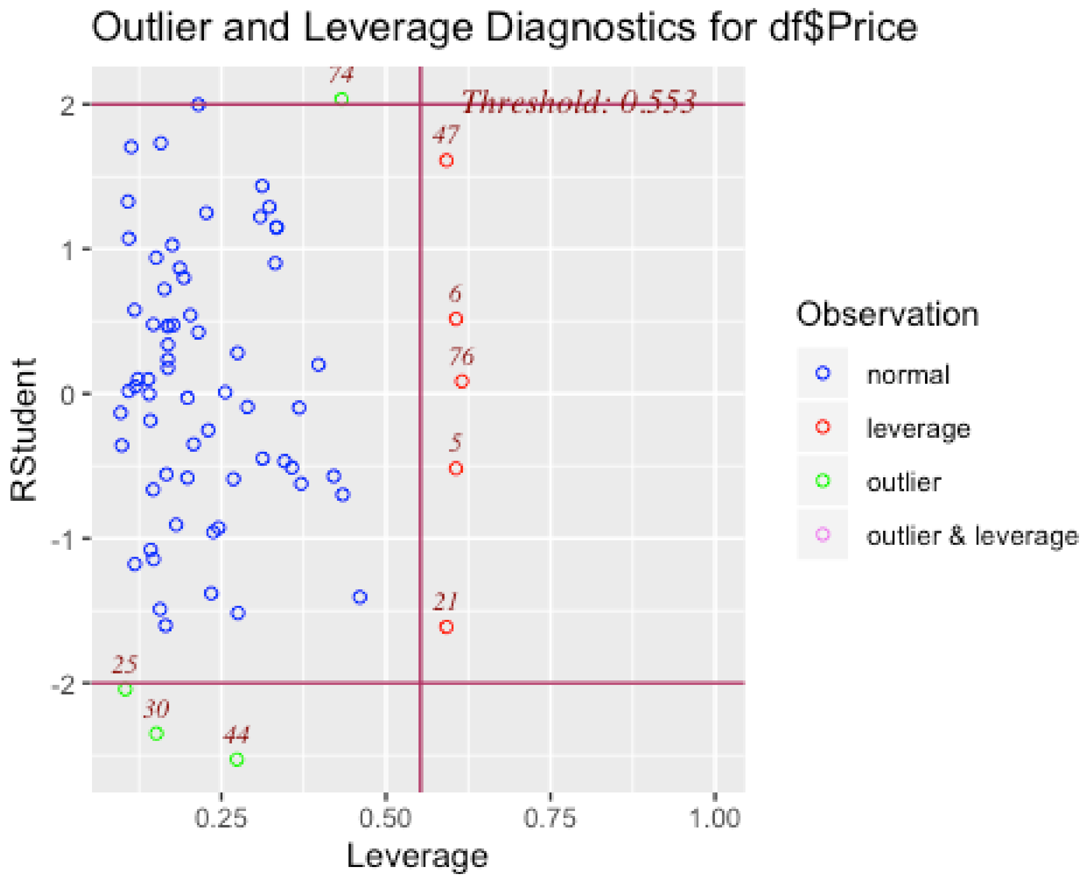
1. If an outlier is due to incorrectly entered or measured data, we should drop the outlier.
2. If the outlier does not change the results but does affect assumptions,then we may drop the outlier.
3. If the outlier affects both results and assumptions. In this situation, it is *not* legitimate to simply drop the outlier. We may run the analysis both with and without it to get clear picture.
4. If the outlier *creates* a significant association, we *should* drop the outlier and *should not* report any significance from your analysis.

```
outlierTest(reg_model)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
## rstudent  unadjusted  p-value  Bonferroni   p
## 44 -2.52766 0.014441       NA
```

- Bonferroni P value is NA and hence there are no significant outliers in the data.

```
ols_plot_resid_lev(reg_model)
```



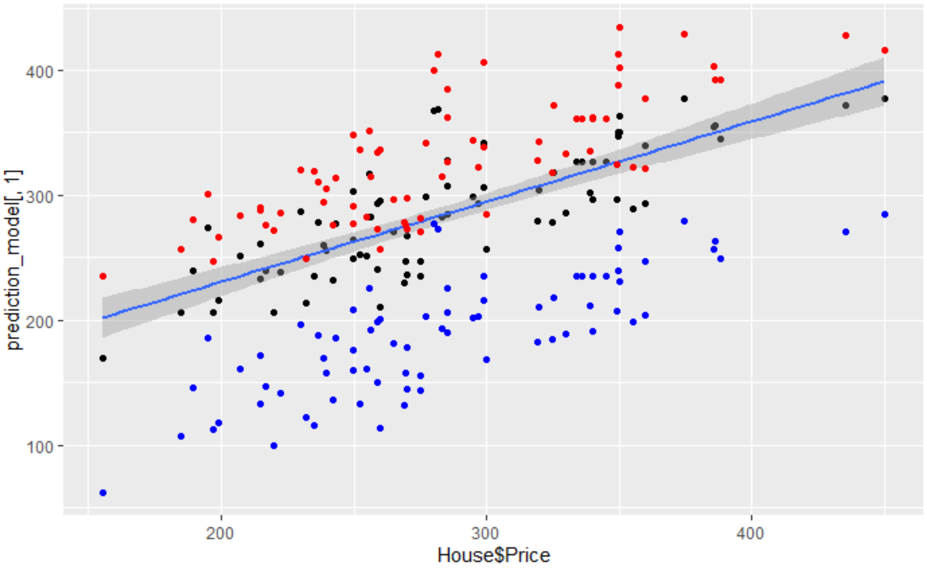
- ols_plot_resid_lev plot provides graph for detecting outliers and/or observations with high leverage.
- In above plot there are five leverage points that are more than threshold value of 0.553. These can be kept in the model as they donot change the model coefficients.
- We can notice four possible outliers in the model but as Bonferroni P value is NA hence, they are not significant but may just need further investigation to confirm.

Expected Value, CI and PI :

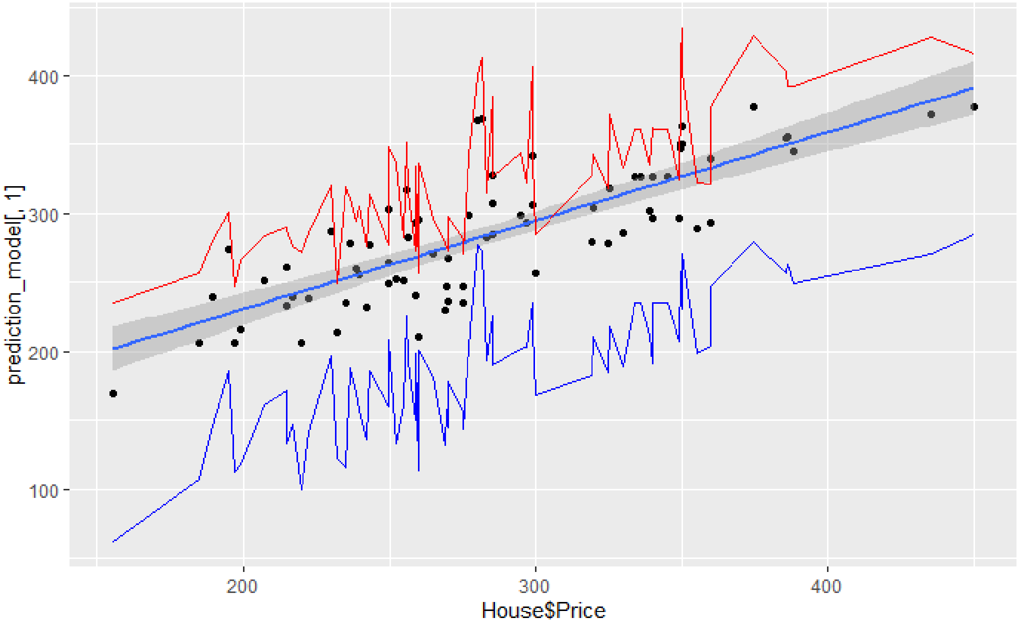
1. Plot the observed house prices, their expected vale (fitted value), confidence intervals (in red) and prediction intervals (in blue). Looking at this plot is this model providing a good estimate of the house prices.

- Confidence intervals tells how well we have determined a parameter of interest, such as a mean or regression coefficient.
- Prediction intervals tells where we can expect to see the next data point sampled.

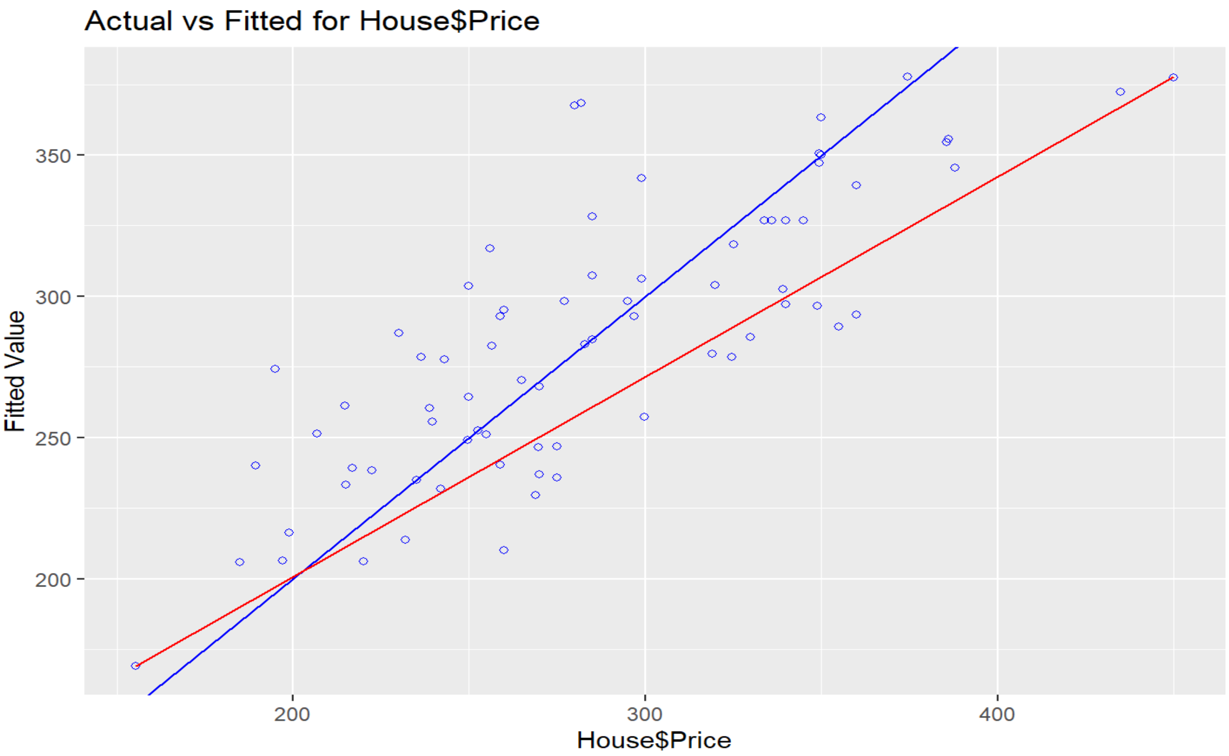
- Plotting Confidence Interval and Prediction Interval with points



- Plotting Confidence Interval and Prediction Interval with lines



```
ols_plot_obs_fit(reg_model)
```



- Plot of observed vs fitted values to assess the fit of the model.
- Ideally, all points should be close to a regressed diagonal line. If model has a high R Square, all the points would be close to this diagonal line. The lower the R Square, the weaker the Goodness of fit of model, the foggier or dispersed your points are from this diagonal line.

Conclusion:

- For our model R square value is 0.625 and the points in the above graph are not scattered far away from the regression line, hence we can say that the model is almost a good fit and provides good estimate of the dependent variable House Price.