House Sale Price Prediction

MSc in Data Analytics January 2021

Sachin Muttappanavar

x20144253@student.ncirl.ie

Abstract: **In this work, relationship between the house characteristics and sale price is interpreted by building model with multiple regression technique.**

I.OBJECTIVES:

- Analyzing data through descriptive statistics and visualization.
- Different multiple regression models are developed. Models are evaluated using appropriate metrics and best model is selected based on its performance.
- Models are verified that Gauss Markov assumptions have been satisfied.
- Summary of final model.

Multiple Linear Regression: Regression model which estimates linear relationship using straight line between continuous dependent variable and more than one independent variables [1].

Equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

β0, β1, . . ., βp: Coefficients of independent variables – It tells how much dependent variable changes with one unit change in independent variable while other variables being constant.

X1, X2..., Xp: Independent variables.

ϵ: Error term

R squared: It is the percentage of variation in dependent variable explained by the linear regression model [1].

R squared value ranges between 0 to 1.

Formula,

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$

Adjusted R squared: Adjust R squared value for the degrees of freedom. It measures percentage of variation explained by independent variables which are significant in predicting dependent variable. Adjusted R squared value decreases when insignificant independent variable is added to model. [1]

$$\bar{R}^2 = 1 - \frac{SS_{\text{res}}/\text{df}_e}{SS_{\text{tot}}/\text{df}_t}$$

Residual Standard Error: Standard deviation of residuals is called s residual Standard Error. If RSE is less, then model is better [1].

Levene Test: Statistical testing like two independent samples T-test and ANOVA test, assume equal variance across groups. Levene test is used for verifying this assumption.

ANOVA hypothesis: ANOVA hypothesis test is conducted to compare means of continuous variable when grouping variable has two or more groups [1].

- Null Hypothesis(H0): all groups have same mean.
- Alternative Hypothesis(H1): one or more sample means are not equal.

Kruskal Test: It is a non-parametric test to compare means between sample groups. If ANOVA test assumptions are not met, Kruskal Test is used as alternative.

Welch's Test: Welch's Test is conducted when levene test of homogeneity of variance has probability value less than 0.05. That is, when variance across sample groups is different [1].

Variance of inflation factor: Variance of inflation factor explains extent of correlation between predictors variable in a model.

If VIF value is,

- 1 - Variable is not correlated with another variable.
- > 4 or > 5 - moderately correlated.
- >= 10 – highly correlated.

Gauss Markov Assumption:

- Linearity: As per gauss markov Linearity assumption there should not be any kind of pattern in relationship between residuals and fitted values if independent variables are linearly related with dependent variables.
- Homoscedasticity: As per gauss markov assumption, residual error variance should be constant in regression. That is, as value of independent variable changes residual error should not vary large.
- Errors are normally distributed.
- Absence of multicollinearity.
- There should not be any influence data points.

II. DESCRIPTION ABOUT DATA SET:

Dataset contains data about various houses in US region.

Dataset dimensions:

Number of columns: 16

Number of observations: 1728

dependent Variable: price

Below snapshot gives information about dataset.

```
> str(house_details)
'data.frame':   1728 obs. of  16 variables:
 $ price       : int  132500 181115 109000 155000 86060 120000 153000 170000 90000 122900 ...
 $ lotSize     : num  0.09 0.92 0.19 0.41 0.11 0.68 0.4 1.21 0.83 1.94 ...
 $ age         : int  42 0 133 13 0 31 33 23 36 4 ...
 $ landvalue   : int  50000 22300 7300 18700 15000 14000 23300 14600 22200 21200 ...
 $ livingarea  : int  906 1953 1944 1944 840 1152 2752 1662 1632 1416 ...
 $ pctcollege  : int  35 51 51 51 51 22 51 35 51 44 ...
 $ bedrooms    : int  2 3 4 3 2 4 4 4 3 3 ...
 $ fireplaces  : int  1 0 1 1 0 1 1 1 0 0 ...
 $ bathrooms   : num  1 2.5 1 1.5 1 1 1.5 1.5 1.5 1.5 ...
 $ rooms       : int  5 6 8 5 3 8 8 9 8 6 ...
 $ heating     : Factor w/ 3 levels "electric","hot air",..: 1 3 3 2 2 2 3 2 1 2 ...
 $ fuel        : Factor w/ 3 levels "electric","gas",..: 1 2 2 2 2 2 3 3 1 2 ...
 $ sewer       : Factor w/ 3 levels "none","public/commercial",..: 3 3 2 3 2 3 3 3 3 1 ...
 $ waterfront  : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ newConstruction: Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 1 1 1 ...
 $ centralAir  : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 1 1 1 ...
```

Variables details:

- price price (US dollars)
- lotSize size of lot (acres)
- age age of house (years)
- landValue value of land (US dollars)
- livingArea living are (square feet)
- pctCollege percent of neighborhood that graduated college
- bedrooms number of bedrooms
- fireplaces number of fireplaces
- bathrooms number of bathrooms (half bathrooms have no shower or tub)
- rooms number of rooms
- heating type of heating system
- fuel fuel used for heating
- sewer type of sewer system
- waterfront whether property includes waterfront
- newConstruction whether the property is a new construction
- centralAir whether the house has central air

Below snapshot shows descriptive statistics of data.

```
> summary(house_details)
     price            lotSize             age          landvalue       livingarea      pctcollege      bedrooms       fireplaces      bathrooms     
 Min.   :  5000   Min.   : 0.0000   Min.   :  0.00   Min.   :  200   Min.   :  616   Min.   :20.00   Min.   :1.000   Min.   :0.0000   Min.   :0.0  
 1st Qu.:145000   1st Qu.: 0.1700   1st Qu.: 13.00   1st Qu.: 15100   1st Qu.:1300   1st Qu.:52.00   1st Qu.:3.000   1st Qu.:0.0000   1st Qu.:1.5  
 Median :189900   Median : 0.3700   Median : 19.00   Median : 25000   Median :1634   Median :57.00   Median :3.000   Median :1.0000   Median :2.0  
 Mean   :211967   Mean   : 0.5002   Mean   : 27.92   Mean   : 34557   Mean   :1755   Mean   :55.57   Mean   :3.155   Mean   :0.6019   Mean   :1.9  
 3rd Qu.:259000   3rd Qu.: 0.5400   3rd Qu.: 34.00   3rd Qu.: 40200   3rd Qu.:2138   3rd Qu.:64.00   3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.:2.5  
 Max.   :775000   Max.   :12.2000   Max.   :225.00   Max.   :412600   Max.   :5228   Max.   :82.00   Max.   :7.000   Max.   :4.0000   Max.   :4.5  
     rooms                heating          fuel                 sewer      waterfront newConstruction centralAir
 Min.   : 2.000   electric     : 305   electric: 315   none           :  12   No :1713   No :1647        No :1093  
 1st Qu.: 5.000   hot air      :1121   gas     :1197   public/commercial:1213   Yes:  15   Yes:  81        Yes: 635  
 Median : 7.000   hot water/steam: 302   oil     : 216   septic         : 503                                        
 Mean   : 7.042                                                                                                     
 3rd Qu.: 8.250                                                                                                     
 Max.   :12.000
```

III. DATA VISUALIZATION:

i. Correlation and Plot:

Correlation is statistical association which explains strength and direction of linear relationship between two variables.

Correlation value ranges from -1 to 1, where

-1 = Negatively correlation

0 = No correlation

1 = Positive correlation

Correlation formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

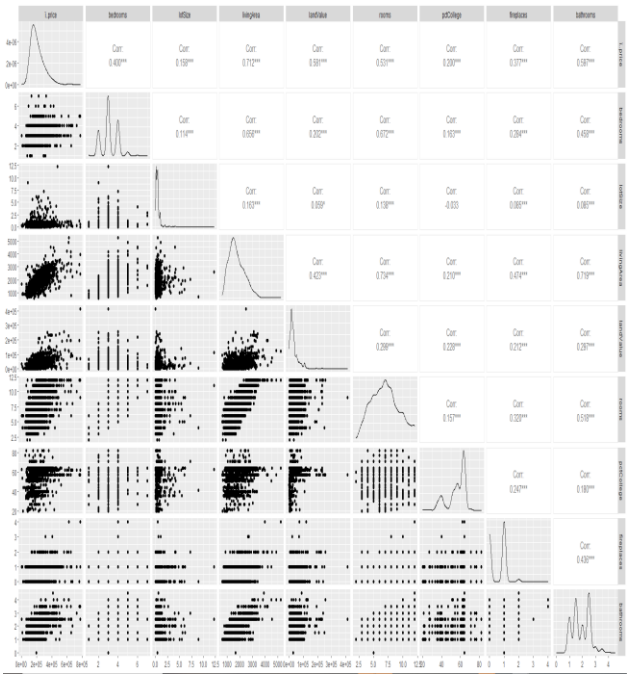
Scatter plots are used to understand correlation. Correlation will be high when data lies on straight line. It reduces as data moves away from the straight line.

If the correlation is high between the independent variables and dependent variable, then we can select those variables in our model for prediction of dependent variable. Correlation value of each numeric independent variable with price variable is as below:

Variables	Price
lotSize	0.158
age	-0.21
landValue	0.581
livingArea	0.712
pctCollege	0.200
bedrooms	0.400
fireplaces	0.377
bathrooms	0.597
rooms	0.531

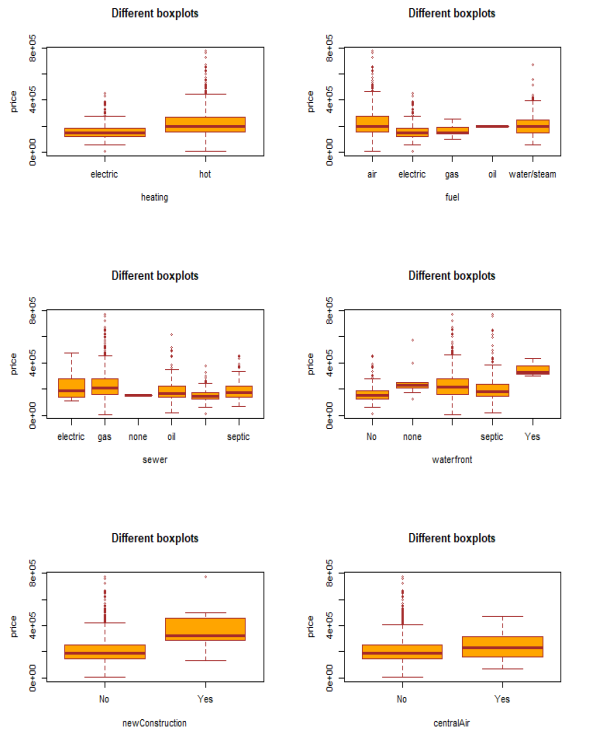
There are no independent variables which have correlation value closer to zero with respect to price. Also, there is no high correlation among the independent variables. Thus, there is no multicollinearity problem in the linear models built with these independent variables.

Below plot depicts correlation value and scatter plot of each variables in dataset.



ii. Box Plot:

Box plot between categorical variables and price.



From box plot, we can see there are some outliers in the price data.

Results of hypothesis test are illustrated below:

Test	Independent Variable	df	p-value	H0
Kruskal-Wallis	heating	2	<2.2e-16	Reject H0
Kruskal-Wallis	Fuel	2	< 2.2e-16	Reject H0
ANOVA	Sewer	2	0.00321	Reject H0
Welch Test	waterfront	14.096	0.001119	Reject H0
ANOVA	newconstruction	1	3.45e-11	Reject H0

As we can see for all the categorical variables, means of dependent variable is not equal between independent groups.

IV. MODELS BUILDING PROCESS AND DESCRIPTION:

As there are outliers in the price data, data is processed to remove observations that are outliers. In total 48 observations are outliers. For analyses of impact of outliers, different models are built with and without outliers on dataset.

Models are compared using following terms:

1) Adjusted R square
2) Residual Stadard Error
3) Probability value

Models with outliers:

Model1: Model is built using numeric independent variables. As we can see all the variables used in model are significant in predicting the house sale price. This model is giving Adjusted R-squared: 0.6221 and Residual standard error : 60510.

```
model1 <- lm(price~livingArea+landValue+bathrooms+rooms+bedrooms, data = house_details)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.700e+04  6.171e+03   2.754  0.00595 **
livingArea   7.169e+01  4.622e+00  15.511  < 2e-16 ***
landValue    9.408e-01  4.620e-02  20.365  < 2e-16 ***
bathrooms    2.691e+04  3.182e+03   8.456  < 2e-16 ***
rooms        2.924e+03  9.967e+02   2.933  0.00340 **
bedrooms    -1.112e+04  2.559e+03  -4.344 1.48e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60510 on 1722 degrees of freedom
Multiple R-squared:  0.6232,    Adjusted R-squared:  0.6221
F-statistic: 569.7 on 5 and 1722 DF,  p-value: < 2.2e-16
```

Second models is built to reduce residual error and improve Adjusted R-squared.

Model2: Model2 is built by adding new independent variables to model1.Newly added variables are newC onstruction, central Air, fuel. The model is giving Ad justed R-Squared:0.6336, Residual error: 59580.

Newly added variables are not making changes to model in predicting dependent variable compared to model1. There is no significant change in Residual standard error and Adjusted R squared value.

```
model2 <- lm(price~bathrooms+lotSize+age+landValue+livingArea+newConstruction+centralAir+fuel, data=house_details)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.550e+03  5.759e+03   1.137 0.255615
bathrooms           2.230e+04  3.389e+03   6.581 6.18e-11 ***
lotSize             7.024e+03  2.170e+03   3.237 0.001232 **
age                -1.936e+02  5.749e+01  -3.368 0.000773 ***
landValue           9.866e-01  4.684e-02  21.065  < 2e-16 ***
livingArea          7.000e+01  3.613e+00  19.372  < 2e-16 ***
newConstructionYes -4.475e+04  7.273e+03  -6.153 9.45e-10 ***
centralAirYes       1.190e+04  3.315e+03   3.591 0.000339 ***
fuelgas             6.754e+03  4.036e+03   1.674 0.094407 .
fueloil             8.293e+03  5.624e+03   1.475 0.140523
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59580 on 1718 degrees of freedom
Multiple R-squared:  0.6355,    Adjusted R-squared:  0.6336
F-statistic: 332.9 on 9 and 1718 DF,  p-value: < 2.2e-16
```

Models with outliers:

Model3: Model is built without outliers in data.

```
model3 <- lm(data = house_details_outliersrm,price~livingArea+landValue+bathrooms+rooms+bedrooms)
```

We can see there is significantly reduce in residual standard error when compared to models with outliers. Thus, outliers are impacting model and resulting in larger residual value. Except rooms all variable are significant in prediction of dependent variable.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 25462.0779  5144.2958   4.950 8.19e-07 ***
livingArea     61.2345     4.0298  15.195  < 2e-16 ***
landValue       0.7649     0.0431  17.748  < 2e-16 ***
bathrooms   26659.5985  2665.5266  10.002  < 2e-16 ***
rooms        1294.7018   828.5832   1.563   0.1183
bedrooms    -4116.2808  2169.2528  -1.898   0.0579 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49570 on 1669 degrees of freedom
Multiple R-squared:  0.6065,    Adjusted R-squared:  0.6053
F-statistic: 514.5 on 5 and 1669 DF,  p-value: < 2.2e-16
```

Model4: Previous model is modified by removing insignificant (p value > 0.05) variable and adding new variables like lotsize, age, newConstruction, centralAir.

```
m <- lm(data = house_details_outliersrm,price~lotSize + age + landValue + livingArea + bedrooms +
  bathrooms  + newConstruction +
  centralAir)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.859e+04  5.381e+03   5.314 1.22e-07 ***
lotSize             8.096e+03  1.899e+03   4.264 2.12e-05 ***
age                -1.759e+02  4.677e+01  -3.761 0.000175 ***
landValue           7.927e-01  4.387e-02  18.071  < 2e-16 ***
livingArea          6.220e+01  3.714e+00  16.749  < 2e-16 ***
bedrooms           -7.057e+02  2.033e+03  -0.347 0.728551
bathrooms           2.173e+04  2.822e+03   7.703 2.27e-14 ***
newConstructionYes -2.386e+04  6.112e+03  -3.903 9.86e-05 ***
centralAirYes       1.103e+04  2.735e+03   4.034 5.72e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 48670 on 1666 degrees of freedom
Multiple R-squared:  0.6213,    Adjusted R-squared:  0.6195
F-statistic: 341.7 on 8 and 1666 DF,  p-value: < 2.2e-16
```

All the variables used in this model are significant in predicting price of house except bedrooms. Prediction of model is not improved from model3. There is no significant reduction in sum of residuals.

Model5: Model5 is built with backward selection to select best subset among independent variables. [2] **Backward selection** is a one of the Best Subset regression technique. These techniques are computationally expensive and greedily search for independent variables which are significant in prediction of dependent variable. Below is code in R for backward selection.

```
model5 <- lm(price ~., data = train_data )
stepbackm = step(model5, direction = "backward")
```

Data is split into train and test data in the ratio 9:1(50+8m – m number of independent variables). Reason for data split is check how model perform with new sample data. [3]

```
set.seed(123)   #  set seed to ensure you always have
sample = sample.split(house_details,SplitRatio = 0.9)
train_data = subset(house_details,sample ==TRUE) # cr
test_data=subset(house_details, sample==FALSE)
train_data <- subset(train_data, select = -c(V17))
test_data <- subset(test_data, select = -c(V17))
```

After running backward selection, it has given below summary of best subset among independent variables.

```
Step:  AIC=33149.59
price ~ lotSize + age + landValue + livingArea + bedrooms + bathrooms +
    rooms + heating + waterfront + newConstruction + centralAir

                   Df  Sum of Sq        RSS   AIC
<none>                           4.9397e+12 33150
- age               1 1.1077e+10 4.9508e+12 33151
- heating           2 2.5353e+10 4.9650e+12 33153
- bedrooms          1 2.1728e+10 4.9614e+12 33154
- rooms             1 2.1938e+10 4.9616e+12 33154
- centralAir        1 2.4668e+10 4.9644e+12 33155
- lotSize           1 3.0133e+10 4.9698e+12 33157
- newConstruction   1 8.7417e+10 5.0271e+12 33174
- bathrooms         1 1.1403e+11 5.0537e+12 33182
- waterfront        1 2.3696e+11 5.1767e+12 33218
- livingArea        1 8.1053e+11 5.7502e+12 33377
- landValue         1 1.0791e+12 6.0188e+12 33446
```

Model5 is modeled based on results from backward selection. Model summary is as below:

```
Call:
lm(formula = price ~ age + heating + bedrooms + rooms + centralAir +
    lotSize + newConstruction + bathrooms + waterfront + livingArea +
    landValue, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-225770  -35349   -4831   27565  398745

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.195e+03  6.957e+03   0.747  0.45533
age                    -1.100e+02  6.002e+01  -1.833  0.06694 .
heatinghot air          8.817e+03  4.291e+03   2.055  0.04007 *
heatinghot water/steam -9.496e+02  5.537e+03  -0.171  0.86386
bedrooms               -6.881e+03  2.680e+03  -2.568  0.01033 *
rooms                   2.613e+03  1.013e+03   2.580  0.00997 **
centralAirYes           9.799e+03  3.582e+03   2.736  0.00629 **
lotSize                 6.639e+03  2.195e+03   3.024  0.00254 **
newConstructionYes     -3.897e+04  7.566e+03  -5.151 2.94e-07 ***
bathrooms               2.082e+04  3.539e+03   5.883 4.97e-09 ***
waterfrontYes           1.382e+05  1.630e+04   8.480  < 2e-16 ***
livingArea              7.448e+01  4.749e+00  15.683  < 2e-16 ***
landValue               8.875e-01  4.905e-02  18.096  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57400 on 1499 degrees of freedom
Multiple R-squared:  0.6595,    Adjusted R-squared:  0.6568
F-statistic: 241.9 on 12 and 1499 DF,  p-value: < 2.2e-16
```

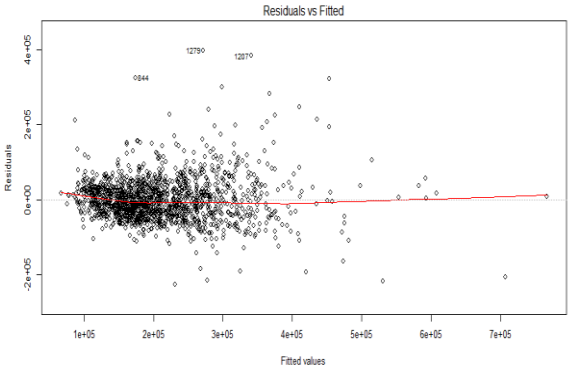
All independent variables used in the model are significant in predicting the dependent variables. Results are almost same as model4 with Residual error of 50830 and adjusted R squared 0.6281.

Below is code to find accuracy of model with training data and testing data.

Accuracy of both training and testing data is almost same. Thus, model is not overfitting the training data.

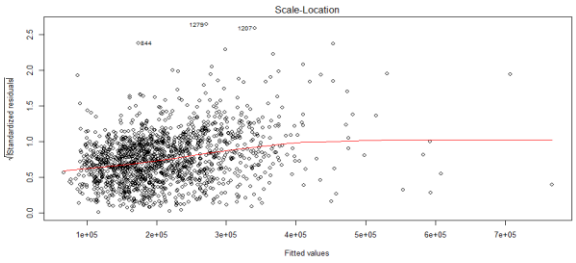
V. DIAGNOSTIC PLOTS FOR MODEL5:

i. Linearity: In the below plot of Residual vs Fitted is not following any systematic pattern. All points are randomly scattered. Hence, linearity assumption is satisfied.



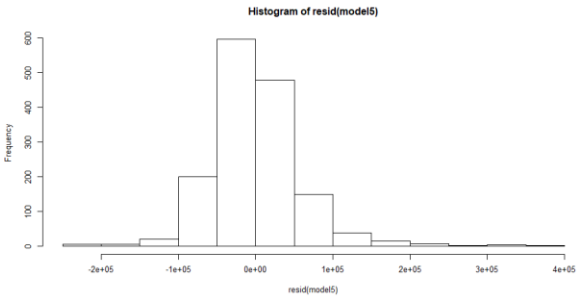
ii. Homscedasticity: There is no systematic pattern in the plot of standardized residuals and Fitted values. All points are randomly scattered. Conducted non constant variance test, results are shown below. It is giving p-value less than 0.05. That is, there is no heteroscedasticity problem in residuals errors. Thus, satisfying homoscedasticity assumption.

```
> plot(model5)
> df_train=data.frame(actual=train_data$price, predicted=model5$fitted.values)
> res_train=mean(abs(df_train$actual-df_train$predicted)/df_train$actual)
> accuracy_train=1-res_train
> accuracy_train
[1] 0.7489609
> df_test=data.frame(actual=test_data$price, predicted=predict(newdata = test_data, model5))
>
> res_test=mean(abs(df_test$actual-df_test$predicted)/df_test$actual)
> accuracy_test=1-res_test
> accuracy_test
[1] 0.7258605
```




```
> ncvTest(model5)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 361.0338, Df = 1, p = < 2.22e-16
```

iii. Normality: Histogram of residuals of model5 follows normal distribution. Thus, normal distribution of residual assumption not violated.

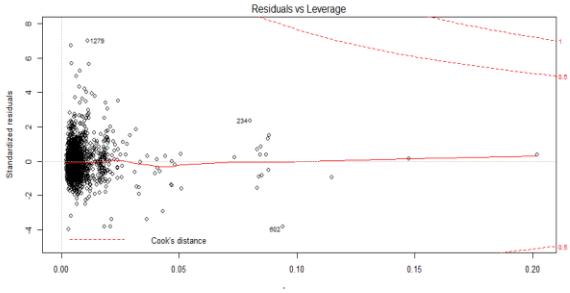


iv. Multicollinearity check: Below is result of variance inflation factor in R for model5. As last column in below figure has values for all variables are in between 1 to 2. Thus, is no problem of multicollinearity.

```
> vif(model5)
                    GVIF Df GVIF^(1/(2*Df))
age             1.398396  1        1.182538
heating         1.409901  2        1.089675
bedrooms        2.170170  1        1.473150
rooms           2.493829  1        1.579186
centralAir      1.374570  1        1.172421
lotSize         1.038921  1        1.019275
newConstruction 1.159678  1        1.076884
bathrooms       2.495538  1        1.579727
waterfront      1.039445  1        1.019532
livingArea      3.930155  1        1.982462
landValue       1.319990  1        1.148908
>
```

iv. Influential Data Point:

To check if there are any influential observations, we look out for Residuals vs Leverage plot. As there are no cases which lie outside dashed red boundary. All cases placed inside cook distance lines. Hence, there are no influential observations.



VI. SUMMARY OF THE FINAL MODEL:

Models are built with outliers are producing high adjusted r squared but models are making big residual errors because of outliers' presence in the data.

Among model3, model4, model5 - model5 is producing good, adjusted R squared value with a smaller number of independent variables. All variables used in models are significant. It is satisfying all the Gauss Markov assumptions. Model5 is considered as final model in this work.

Below is the final multiple regression model formula and summary:

```
Call:
lm(formula = price ~ age + heating + bedrooms + rooms + centralAir +
    lotSize + newConstruction + bathrooms + waterfront + livingArea +
    landValue, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-225770  -35349   -4831   27565  398745

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.195e+03  6.957e+03   0.747  0.45533
age                    -1.100e+02  6.002e+01  -1.833  0.06694 .
heatinghot air          8.817e+03  4.291e+03   2.055  0.04007 *
heatinghot water/steam -9.496e+02  5.537e+03  -0.171  0.86386
bedrooms               -6.881e+03  2.680e+03  -2.568  0.01033 *
rooms                   2.613e+03  1.013e+03   2.580  0.00997 **
centralAirYes           9.799e+03  3.582e+03   2.736  0.00629 **
lotSize                 6.639e+03  2.195e+03   3.024  0.00254 **
newConstructionYes     -3.897e+04  7.566e+03  -5.151 2.94e-07 ***
bathrooms               2.082e+04  3.539e+03   5.883 4.97e-09 ***
waterfrontYes           1.382e+05  1.630e+04   8.480  < 2e-16 ***
livingArea              7.448e+01  4.749e+00  15.683  < 2e-16 ***
landValue               8.875e-01  4.905e-02  18.096  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57400 on 1499 degrees of freedom
Multiple R-squared:  0.6595,    Adjusted R-squared:  0.6568
F-statistic: 241.9 on 12 and 1499 DF,  p-value: < 2.2e-16
```

Model5 accuracy in training and testing data is 0.75 and 0.72 respectively which are almost similar. Thus, Model5 is not overfitting training data and model is generalized for out of sample data.

Model can be improved by penalized regression methods such as ridge regression, lasso regression, and elastic net regression.

REFERENCES:

[1]. Practical Statistic for Data Scientist by orielly book.

[2]. Feature engineering for machine learning by orielly book.

[3]. Introduction to statistical Learning with application in R.

[4]. Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning - IEEE Paper.

[5]. Jeremy J Foster,Emma Barkus,Christian Yavorsky. (2006), Understanding and Using Advanced Statistics, SAGE, p.178, [ISBN: 141290014X].