Business Analytics – Term Project

Zillow's Zestimate
Home Valuation

BY

Pusarapu Rajendra

Gujja Rithin Rao

Mandumula Sai Abhinav

Tandulwadkar Sharmili

Under the Guidance

Of

ROUZBEH RAZAVI, PH.D.

Contributions

Everyone actively participated as a member of a team to move the team toward the completion of goals. By concerning individual area of expertise to assist the team to complete the project.

Pusarapu Rajendra	Data-Prep, Scripting, Document
Gujja Rithin Rao	Data- Prep, Scripting, Graph Predictions
Mandumula Sai Abhinav	Strategy, Scripting, Documentation, PPT
Tandulwadkar Sharmili	Documentation, PPT

Contents

Introduction

Since its first publication 11 years ago, Zillow's Zestimate home valuation has shaken the U.S. real estate industry. A home is often a person's largest and most expensive purchase in life. It is extremely important to ensure that homeowners have a reliable way to monitor this property. The Zestimate was developed to provide customers with as much information as possible about homes and the housing market, marking the first-time consumers had access to information about this type of home value at no cost.

"Zestimates" are approximate home values based on 7.5 million computer and mathematical modeling systems which evaluate hundreds of data points on each property. And by continuously improving the median error margin (from 14% at the beginning to 5% today), Since then Zillow has become one of the best, most respected real estate data marketplaces in the United States and a pioneering instance of impactful machine learning.

Project Goal

The key goal is to improve the estimation of housing price of buyers in the real estate industry by the Zillow service.

Data Exploration

A. Overview of Data

The train data consist of 1000 observations for which sales price of the house can be predicted from 36 variables.

Summary:

```
       Id           MSSubClass      LotFrontage       LotArea        OverallQual

 Min.   :   1.0   Min.   : 20.00   Min.   : 21.00   Min.   :  1300   Min.   : 1.000
 1st Qu.: 250.8   1st Qu.: 20.00   1st Qu.: 60.00   1st Qu.:  7585   1st Qu.: 5.000
 Median : 500.5   Median : 50.00   Median : 70.00   Median :  9451   Median : 6.000
 Mean   : 500.5   Mean   : 56.88   Mean   : 69.96   Mean   : 10691   Mean   : 6.125
```

```
3rd Qu.: 750.2   3rd Qu.: 70.00   3rd Qu.: 80.00   3rd Qu.: 11628   3rd Qu.: 7.000
Max.   :1000.0   Max.   :190.00   Max.   :313.00   Max.   :215245   Max.   :10.000
                                  NA's   :173
 OverallCond      YearBuilt     YearRemodAdd    MasVnrArea       BsmtFinSF1
Min.   :1.000   Min.   :1880   Min.   :1950   Min.   :   0.0   Min.   :   0.0
1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   1st Qu.:   0.0   1st Qu.:   0.0
Median :5.000   Median :1974   Median :1994   Median :   0.0   Median : 384.5
Mean   :5.587   Mean   :1972   Mean   :1985   Mean   : 109.2   Mean   : 445.2
3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004   3rd Qu.: 174.8   3rd Qu.: 725.0
Max.   :9.000   Max.   :2010   Max.   :2010   Max.   :1600.0   Max.   :2260.0
                                              NA's   :6
  BsmtFinSF2        BsmtUnfSF        X1stFlrSF        X2ndFlrSF       LowQualFinSF
Min.   :   0.0   Min.   :   0.0   Min.   : 334.0   Min.   :   0.0   Min.   :  0.000
1st Qu.:   0.0   1st Qu.: 226.5   1st Qu.: 876.8   1st Qu.:   0.0   1st Qu.:  0.000
Median :   0.0   Median : 474.0   Median :1087.0   Median :   0.0   Median :  0.000
Mean   :  48.3   Mean   : 567.8   Mean   :1157.0   Mean   : 347.1   Mean   :  6.474
3rd Qu.:   0.0   3rd Qu.: 808.0   3rd Qu.:1389.5   3rd Qu.: 728.2   3rd Qu.:  0.000
Max.   :1474.0   Max.   :2336.0   Max.   :3228.0   Max.   :1872.0   Max.   :572.000

 BsmtFullBath    BsmtHalfBath      FullBath        HalfBath       BedroomAbvGr    KitchenAbvGr
Min.   :0.00   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
1st Qu.:0.00   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:2.000   1st Qu.:1.000
Median :0.00   Median :0.000   Median :2.000   Median :0.000   Median :3.000   Median :1.000
Mean   :0.43   Mean   :0.058   Mean   :1.566   Mean   :0.386   Mean   :2.855   Mean   :1.047
3rd Qu.:1.00   3rd Qu.:0.000   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.:1.000
Max.   :3.00   Max.   :2.000   Max.   :3.000   Max.   :2.000   Max.   :8.000   Max.   :3.000

 TotRmsAbvGrd      Fireplaces     GarageYrBlt     GarageCars      GarageArea
Min.   : 2.000   Min.   :0.00   Min.   :1900   Min.   :0.000   Min.   :   0.0
1st Qu.: 5.000   1st Qu.:0.00   1st Qu.:1961   1st Qu.:1.000   1st Qu.: 338.0
Median : 6.000   Median :1.00   Median :1980   Median :2.000   Median : 480.0
Mean   : 6.495   Mean   :0.61   Mean   :1979   Mean   :1.765   Mean   : 473.4
3rd Qu.: 7.000   3rd Qu.:1.00   3rd Qu.:2002   3rd Qu.:2.000   3rd Qu.: 576.0
Max.   :14.000   Max.   :3.00   Max.   :2010   Max.   :4.000   Max.   :1390.0
                                NA's   :56
  WoodDeckSF      OpenPorchSF     EnclosedPorch      X3SsnPorch       ScreenPorch
Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Min.   :  0.00
1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.00
Median :  0.00   Median : 24.00   Median :  0.00   Median :  0.000   Median :  0.00
Mean   : 97.35   Mean   : 47.67   Mean   : 21.41   Mean   :  3.703   Mean   : 15.05
3rd Qu.:171.25   3rd Qu.: 70.00   3rd Qu.:  0.00   3rd Qu.:  0.000   3rd Qu.:  0.00
Max.   :857.00   Max.   :523.00   Max.   :552.00   Max.   :508.000   Max.   :410.00

   PoolArea          MiscVal             MoSold           YrSold       SalePrice
Min.   :  0.00   Min.   :    0.00   Min.   : 1.000   Min.   :2006   Min.   : 34900
1st Qu.:  0.00   1st Qu.:    0.00   1st Qu.: 5.000   1st Qu.:2007   1st Qu.:130000
Median :  0.00   Median :    0.00   Median : 6.000   Median :2008   Median :163995
Mean   :  1.16   Mean   :   45.38   Mean   : 6.307   Mean   :2008   Mean   :182285
3rd Qu.:  0.00   3rd Qu.:    0.00   3rd Qu.: 8.000   3rd Qu.:2009   3rd Qu.:215000
Max.   :648.00   Max.   :15500.00   Max.   :12.000   Max.   :2010   Max.   :755000
```

B. Data Cleaning and Preparation

We found the data has missing values. This data is to be treated before feeding to the algorithm because missing values could distort the model performance.

The variables which have missing values are “Lot Frontage”, ”MasVnrArea”,”GarageYrBlt”.

The missing values are predicted with most appropriate method “KNN algorithm”.

In KNN method we find the optimal k value by “search grid” method. Hence NA values are filled by most appropriate value predicted by using this method.

We are eliminating the first variable (member id) because it doesn’t contribute significantly to predict the outcome.

Summary :

```
MSSubClass       LotFrontage        LotArea         OverallQual      OverallCond
 Min.   : 20.00   Min.   : 21.00   Min.   :   1300   Min.   : 1.000   Min.   :1.000
 1st Qu.: 20.00   1st Qu.: 60.00   1st Qu.:   7585   1st Qu.: 5.000   1st Qu.:5.000
 Median : 50.00   Median : 70.00   Median :   9451   Median : 6.000   Median :5.000
 Mean   : 56.88   Mean   : 69.67   Mean   : 10691   Mean   : 6.125   Mean   :5.587
 3rd Qu.: 70.00   3rd Qu.: 80.00   3rd Qu.: 11628   3rd Qu.: 7.000   3rd Qu.:6.000
 Max.   :190.00   Max.   :313.00   Max.   :215245   Max.   :10.000   Max.   :9.000
   YearBuilt     YearRemodAdd      MasVnrArea        BsmtFinSF1       BsmtFinSF2
 Min.   :1880   Min.   :1950   Min.   :   0.0   Min.   :   0.0   Min.   :   0.0
 1st Qu.:1954   1st Qu.:1967   1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0
 Median :1974   Median :1994   Median :   0.0   Median : 384.5   Median :   0.0
 Mean   :1972   Mean   :1985   Mean   : 108.9   Mean   : 445.2   Mean   :  48.3
 3rd Qu.:2000   3rd Qu.:2004   3rd Qu.: 174.0   3rd Qu.: 725.0   3rd Qu.:   0.0
 Max.   :2010   Max.   :2010   Max.   :1600.0   Max.   :2260.0   Max.   :1474.0
   BsmtUnfSF        X1stFlrSF        X2ndFlrSF       LowQualFinSF      BsmtFullBath
 Min.   :   0.0   Min.   : 334.0   Min.   :   0.0   Min.   :  0.000   Min.   :0.00
 1st Qu.: 226.5   1st Qu.: 876.8   1st Qu.:   0.0   1st Qu.:  0.000   1st Qu.:0.00
 Median : 474.0   Median :1087.0   Median :   0.0   Median :  0.000   Median :0.00
 Mean   : 567.8   Mean   :1157.0   Mean   : 347.1   Mean   :  6.474   Mean   :0.43
 3rd Qu.: 808.0   3rd Qu.:1389.5   3rd Qu.: 728.2   3rd Qu.:  0.000   3rd Qu.:1.00
 Max.   :2336.0   Max.   :3228.0   Max.   :1872.0   Max.   :572.000   Max.   :3.00
  BsmtHalfBath      FullBath        HalfBath       BedroomAbvGr     KitchenAbvGr
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
 1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:2.000   1st Qu.:1.000
 Median :0.000   Median :2.000   Median :0.000   Median :3.000   Median :1.000
 Mean   :0.058   Mean   :1.566   Mean   :0.386   Mean   :2.855   Mean   :1.047
 3rd Qu.:0.000   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.:1.000
 Max.   :2.000   Max.   :3.000   Max.   :2.000   Max.   :8.000   Max.   :3.000
  TotRmsAbvGrd      Fireplaces      GarageYrBlt      GarageCars      GarageArea
 Min.   : 2.000   Min.   :0.00   Min.   :1900   Min.   :0.000   Min.   :   0.0
 1st Qu.: 5.000   1st Qu.:0.00   1st Qu.:1959   1st Qu.:1.000   1st Qu.: 338.0
 Median : 6.000   Median :1.00   Median :1978   Median :2.000   Median : 480.0
 Mean   : 6.495   Mean   :0.61   Mean   :1977   Mean   :1.765   Mean   : 473.4
 3rd Qu.: 7.000   3rd Qu.:1.00   3rd Qu.:2001   3rd Qu.:2.000   3rd Qu.: 576.0
 Max.   :14.000   Max.   :3.00   Max.   :2010   Max.   :4.000   Max.   :1390.0
   WoodDeckSF      OpenPorchSF     EnclosedPorch       X3SsnPorch       ScreenPorch
 Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Min.   :  0.00
 1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.00
 Median :  0.00   Median : 24.00   Median :  0.00   Median :  0.000   Median :  0.00
 Mean   : 97.35   Mean   : 47.67   Mean   : 21.41   Mean   :  3.703   Mean   : 15.05
 3rd Qu.:171.25   3rd Qu.: 70.00   3rd Qu.:  0.00   3rd Qu.:  0.000   3rd Qu.:  0.00
 Max.   :857.00   Max.   :523.00   Max.   :552.00   Max.   :508.000   Max.   :410.00
    PoolArea         MiscVal               MoSold           YrSold       SalePrice
 Min.   :  0.00   Min.   :    0.00   Min.   : 1.000   Min.   :2006   Min.   : 34900
 1st Qu.:  0.00   1st Qu.:    0.00   1st Qu.: 5.000   1st Qu.:2007   1st Qu.:130000
 Median :  0.00   Median :    0.00   Median : 6.000   Median :2008   Median :163995
 Mean   :  1.16   Mean   :   45.38   Mean   : 6.307   Mean   :2008   Mean   :182285
 3rd Qu.:  0.00   3rd Qu.:    0.00   3rd Qu.: 8.000   3rd Qu.:2009   3rd Qu.:215000
 Max.   :648.00   Max.   :15500.00   Max.   :12.000   Max.   :2010   Max.   :755000
```

Modelling Strategy

Regression analysis is a method for numerical, predictive modeling used to analyze the association between a dependent variable and one or more independent variables.

The house details, such as the Dwelling, Zone of Site, Material, etc., are referred to as the independent or predictor variables. Such parameter predictors are used to predict the variable response. For our situation, the solution parameter is the house price. Answer variables are also referred to as dependent variables because their values depend on the independent variable values.

So, Provided the house's relevant data, our job is to predict a future cost.

There are many methods for regression analysis, but the most widely applied and regression model is Linear Regression.

In a linear regression method, the interaction between the dependent and independent variables is always continuous, so you will see more of a straight line than a curved line as you decide to map their relationship.

As we have more dependent variables towards a single dependent variable we choose multiple regression which is an extension to linear regression .It also allows you to determine the overall fit (variance explained) of the model and the relative contribution of each of the predictors to the total variance explained.

Model's Performance

As We choose Multiple regression to build out model our initial step was to tune the data by eliminating the NA values and replacing with proper justifications as said above.

```
k_na<-na.omit(train_1_) # Removing the NA Rows.

# Predicting the K Value
search_grid<-expand.grid(k=c(1:20))
K<-train(SalePrice ~.,data=k_na,method="knn",tuneGrid=search_grid,preProcess='range')
```

```r
# Imputing NA Values by KNN Method
train1<-kNN(train_1_,variable=c("LotFrontage","MasVnrArea","GarageYrBlt"),k=13)
train1<-train1[,-c(1,37,38,39)]
```

Code – Replacing the NA values with appropriate.

Now that we have our Data ready for model building, we start building our model on our train data set.

Model 1:

```r
model1<-lm(SalePrice~.,data = train1)
summary(model1) #Summary of the model
```

##

Call:

lm(formula = SalePrice ~ ., data = train1)

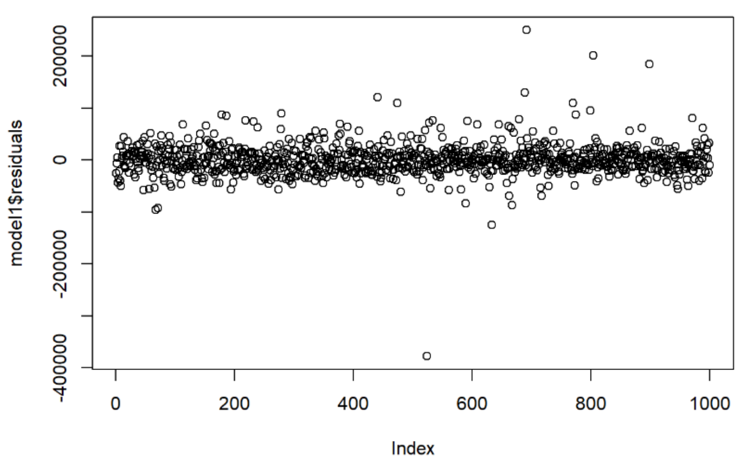
##

Residuals:

Min 1Q Median 3Q Max

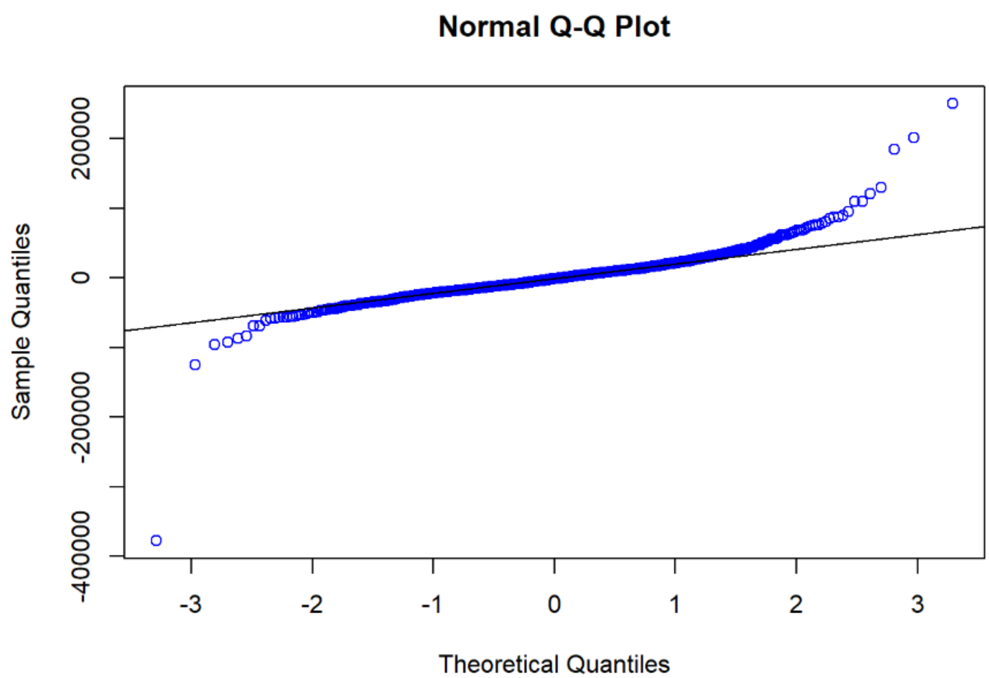
```
##
## Call:
## lm(formula = SalePrice ~ ., data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -377759  -15646   -1228   12756  249812
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.776e+05  1.565e+06   0.497 0.619351
## MSSubClass    -1.113e+02  3.113e+01  -3.574 0.000369 ***
## LotFrontage    7.566e+01  6.049e+01   1.251 0.211281
## LotArea        4.527e-01  9.822e-02   4.609 4.58e-06 ***
## OverallQual    1.560e+04  1.348e+03  11.577  < 2e-16 ***
## OverallCond    3.198e+03  1.170e+03   2.732 0.006402 **
## YearBuilt      2.715e+02  7.898e+01   3.438 0.000611 ***
## YearRemodAdd   2.829e+02  7.569e+01   3.737 0.000197 ***
## MasVnrArea     2.915e+01  6.342e+00   4.596 4.89e-06 ***
## BsmtFinSF1     3.506e+01  5.431e+00   6.456 1.70e-10 ***
## BsmtFinSF2     1.963e+01  7.686e+00   2.554 0.010816 *
## BsmtUnfSF      1.887e+01  4.857e+00   3.886 0.000109 ***
## X1stFlrSF      4.776e+01  6.599e+00   7.237 9.36e-13 ***
## X2ndFlrSF      4.748e+01  5.523e+00   8.596  < 2e-16 ***
## LowQualFinSF   4.122e+01  2.094e+01   1.969 0.049281 *
## BsmtFullBath   4.494e+03  2.916e+03   1.541 0.123604
## BsmtHalfBath   9.205e+02  4.527e+03   0.203 0.838913
## FullBath       1.722e+03  3.158e+03   0.545 0.585782
## HalfBath       3.990e+02  2.975e+03   0.134 0.893324
## BedroomAbvGr  -1.366e+04  1.885e+03  -7.246 8.77e-13 ***
## KitchenAbvGr  -1.794e+04  5.679e+03  -3.159 0.001633 **
## TotRmsAbvGrd   8.107e+03  1.393e+03   5.822 7.93e-09 ***
## Fireplaces     3.145e+03  1.942e+03   1.620 0.105612
## GarageYrBlt    4.616e+01  8.652e+01   0.533 0.593811
## GarageCars    -5.910e+02  3.188e+03  -0.185 0.852971
## GarageArea     2.948e+01  1.129e+01   2.611 0.009167 **
## WoodDeckSF     1.545e+01  8.934e+00   1.729 0.084150 .
## OpenPorchSF   -1.184e+01  1.650e+01  -0.717 0.473254
## EnclosedPorch  6.099e+00  1.903e+01   0.321 0.748585
## X3SsnPorch    -2.693e+01  3.277e+01  -0.822 0.411304
## ScreenPorch    5.578e+01  1.902e+01   2.933 0.003440 **
## PoolArea      -5.428e+01  4.045e+01  -1.342 0.179885
## MiscVal       -1.733e-01  1.926e+00  -0.090 0.928331
## MoSold        -6.678e+02  3.848e+02  -1.736 0.082946 .
## YrSold        -1.004e+03  7.812e+02  -1.285 0.199154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31640 on 965 degrees of freedom
## Multiple R-squared:   0.85,  Adjusted R-squared:  0.8447
## F-statistic: 160.8 on 34 and 965 DF,  p-value: < 2.2e-16
```

```
#R-Square value 85 implies that it captured  85% of Variability
plot(model1$residuals) # Constant Variance
```



From the model we can say that its 85% accurate based on the R-Square Value with entire train data variables. The p-Value of the F-Statistic says that at least one variable in the model is statistically significant. By the PR value we can say that there are 18 variable which are statistically significant when we can negotiate the null hypotheses. Residual standard error: 31640

While plotting Residuals of the model it is having constant variance and normally distributed so that we can fit our suggested model.



Residual Plot vs Residual Line Chart

To find the variable importance of the model we apply anova and obtained following:

```
#Variable importance of the model
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: SalePrice
##                Df     Sum Sq    Mean Sq   F value    Pr(>F)
## MSSubClass      1 4.2518e+10 4.2518e+10   42.4604 1.160e-10 ***
## LotFrontage     1 8.5453e+11 8.5453e+11  853.3817 < 2.2e-16 ***
## LotArea         1 1.7420e+11 1.7420e+11  173.9676 < 2.2e-16 ***
## OverallQual     1 3.3767e+12 3.3767e+12 3372.1049 < 2.2e-16 ***
## OverallCond     1 3.0552e+09 3.0552e+09    3.0511  0.080999 .
## YearBuilt       1 6.8258e+10 6.8258e+10   68.1665 4.909e-16 ***
## YearRemodAdd    1 3.5754e+10 3.5754e+10   35.7060 3.225e-09 ***
## MasVnrArea      1 1.5756e+11 1.5756e+11  157.3449 < 2.2e-16 ***
## BsmtFinSF1      1 1.6019e+11 1.6019e+11  159.9736 < 2.2e-16 ***
## BsmtFinSF2      1 6.8087e+09 6.8087e+09    6.7995  0.009259 **
## BsmtUnfSF       1 6.2397e+10 6.2397e+10   62.3128 7.930e-15 ***
## X1stFlrSF       1 5.1403e+10 5.1403e+10   51.3342 1.546e-12 ***
## X2ndFlrSF       1 3.4771e+11 3.4771e+11  347.2445 < 2.2e-16 ***
## LowQualFinSF    1 5.6137e+09 5.6137e+09    5.6061  0.018094 *
## BsmtFullBath    1 2.7686e+09 2.7686e+09    2.7649  0.096680 .
## BsmtHalfBath    1 2.2836e+08 2.2836e+08    0.2281  0.633080
## FullBath        1 1.8189e+09 1.8189e+09    1.8164  0.178053
## HalfBath        1 2.6116e+08 2.6116e+08    0.2608  0.609684
## BedroomAbvGr    1 4.0309e+10 4.0309e+10   40.2547 3.422e-10 ***
## KitchenAbvGr    1 3.0719e+09 3.0719e+09    3.0678  0.080175 .
## TotRmsAbvGrd    1 3.8985e+10 3.8985e+10   38.9330 6.556e-10 ***
## Fireplaces      1 2.2591e+09 2.2591e+09    2.2560  0.133422
## GarageYrBlt     1 5.2468e+09 5.2468e+09    5.2397  0.022292 *
## GarageCars      1 8.9245e+09 8.9245e+09    8.9125  0.002904 **
## GarageArea      1 6.2713e+09 6.2713e+09    6.2628  0.012494 *
## WoodDeckSF      1 2.0046e+09 2.0046e+09    2.0019  0.157421
## OpenPorchSF     1 4.9392e+08 4.9392e+08    0.4933  0.482647
## EnclosedPorch   1 4.2937e+07 4.2937e+07    0.0429  0.835997
## X3SsnPorch      1 9.2696e+08 9.2696e+08    0.9257  0.336222
## ScreenPorch     1 8.3411e+09 8.3411e+09    8.3299  0.003987 **
## PoolArea        1 1.2547e+09 1.2547e+09    1.2530  0.263255
## MiscVal         1 4.5196e+06 4.5196e+06    0.0045  0.946450
## MoSold          1 2.4109e+09 2.4109e+09    2.4076  0.121072
## YrSold          1 1.6531e+09 1.6531e+09    1.6508  0.199154
## Residuals     965 9.6630e+11 1.0013e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

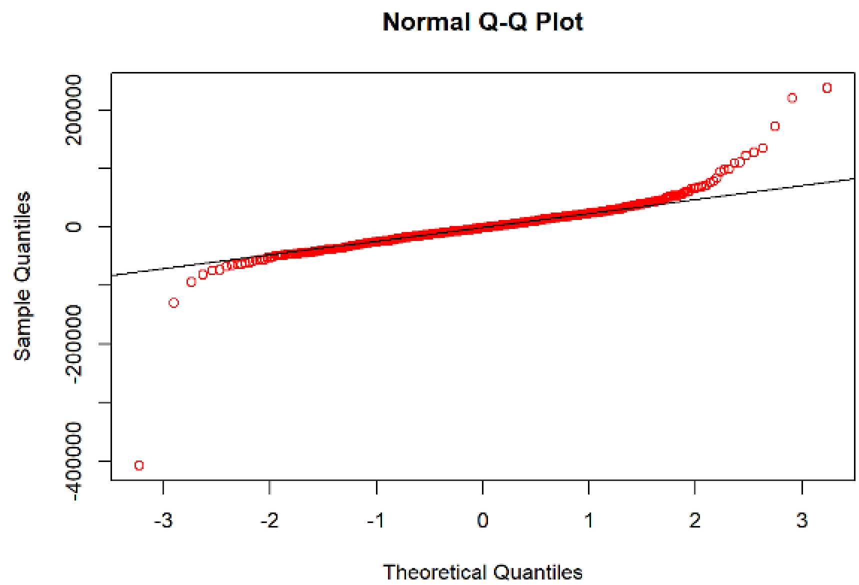
Model 2:

By considering the variables which are statistically significant by taking the PR value cutoff as less than 0.000001.

```
##
## Call:
## lm(formula = SalePrice ~ ., data = model2_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -408479  -16103   -1062   15905  237498
##
## Coefficients:
##                   Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)  -1330329.4731  142798.3234  -9.316 < 0.0000000000000002 ***
## MSSubClass       -170.3026      32.3126  -5.270       0.000000174685 ***
## LotFrontage        55.4752      65.2249   0.851               0.3953
## LotArea             0.6994       0.1498   4.670       0.000003528785 ***
## OverallQual     20126.2774    1403.2762  14.342 < 0.0000000000000002 ***
## YearBuilt         270.5831      52.2140   5.182       0.000000277147 ***
## YearRemodAdd      372.8025      75.3826   4.945       0.000000923952 ***
## MasVnrArea         36.8752       7.2797   5.065       0.000000504904 ***
## BsmtFinSF1         25.6570       3.0526   8.405 < 0.0000000000000002 ***
## BsmtFinSF2         14.8382       7.7605   1.912               0.0562 .
## X1stFlrSF          64.8779       6.1721  10.512 < 0.0000000000000002 ***
## X2ndFlrSF          53.4498       5.2100  10.259 < 0.0000000000000002 ***
## BedroomAbvGr   -13429.0872    2107.3853  -6.372       0.000000000312 ***
## TotRmsAbvGrd     6729.3344    1527.5642   4.405       0.000011986134 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34110 on 809 degrees of freedom
##   (177 observations deleted due to missingness)
## Multiple R-squared:  0.8421, Adjusted R-squared:  0.8396
## F-statistic: 331.9 on 13 and 809 DF,  p-value: < 0.00000000000000022
```

From the model we can say that its 84.21% accurate based on the R-Square Value with entire train data variables. The p-Value of the F-Statistic says that at least one variable in the model is statistically significant. By the PR value we can say that there are 13 variable which are statistically significant when we can negotiate the null hypotheses. Residual standard error: 34110

```r
qqnorm(model2$residuals,col='red') ## Residual plot
qqline(model2$residuals) ## Residual line
```



Residual Plot vs Residual Line Chart

Variable Importance:

```
## Analysis of Variance Table
##
## Response: SalePrice
##                Df        Sum Sq       Mean Sq  F value
## MSSubClass      1   60104013400   60104013400   51.658
## LotFrontage     1  781913968163  781913968163  672.042
## LotArea         1  209676962873  209676962873  180.214
## OverallQual     1 3139931834321 3139931834321 2698.718
## YearBuilt       1   59989935432   59989935432   51.560
## YearRemodAdd    1   37180930942   37180930942   31.956
## MasVnrArea      1  172783345318  172783345318  148.504
## BsmtFinSF1      1  140992116346  140992116346  121.180
## BsmtFinSF2      1    2021019950    2021019950    1.737
## X1stFlrSF       1   83662944909   83662944909   71.907
## X2ndFlrSF       1  282022210246  282022210246  242.393
## BedroomAbvGr    1   27546983325   27546983325   23.676
## TotRmsAbvGrd    1   22579166278   22579166278   19.406
## Residuals     809  941263363585    1163489943
##                               Pr(>F)
## MSSubClass        0.000000000001502 ***
## LotFrontage  < 0.0000000000000022 ***
## LotArea      < 0.0000000000000022 ***
## OverallQual  < 0.0000000000000022 ***
## YearBuilt         0.000000000001575 ***
## YearRemodAdd      0.000000021858076 ***
## MasVnrArea   < 0.0000000000000022 ***
## BsmtFinSF1   < 0.0000000000000022 ***
## BsmtFinSF2                   0.1879
## X1stFlrSF    < 0.0000000000000022 ***
## X2ndFlrSF    < 0.0000000000000022 ***
## BedroomAbvGr      0.000001370275848 ***
## TotRmsAbvGrd      0.000011986134170 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

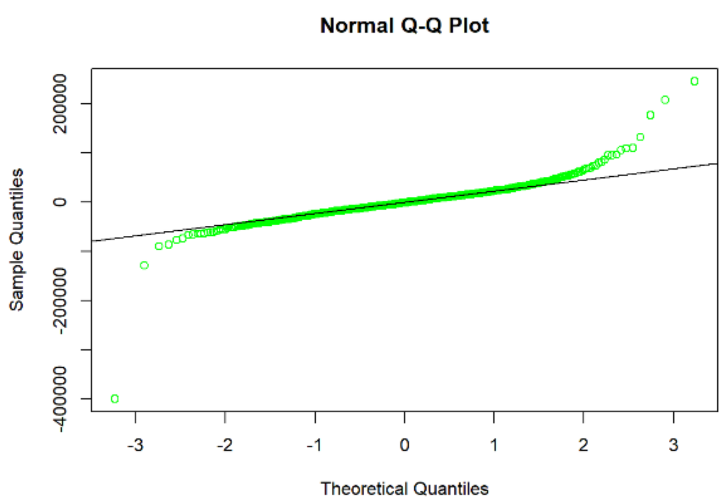
Model 3:

Like the above approaches, achieved 84.75 % accuracy with 16 variables

```r
#By considering the variables which are statistically significant by  taking the pr value cutoff as less than 0.01
# Building the model
model3_data<-train_1_[,c(2,3,4,5,7,8,9,10,11,12,13,14,20,22,25,31,36)]
model3<-lm(SalePrice~.,data = model3_data)
summary(model3) ## Taking 16 variables
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = model3_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -399903  -15769    -881   14832  245381
##
## Coefficients:
##                  Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)  -1265432.7922  144533.2935  -8.755 < 0.0000000000000002 ***
## MSSubClass       -151.0028      32.1483  -4.697     0.00000310187174 ***
## LotFrontage        48.3676      64.7477   0.747              0.45527
## LotArea             0.6822       0.1477   4.619     0.00000448547165 ***
## OverallQual     17781.2840    1472.8771  12.072 < 0.0000000000000002 ***
## YearBuilt         224.5868      54.1042   4.151     0.00003662585324 ***
## YearRemodAdd      386.9553      74.3708   5.203     0.00000024887315 ***
## MasVnrArea         34.4339       7.1923   4.788     0.00000200776361 ***
## BsmtFinSF1         39.5537       5.6267   7.030     0.00000000000441 ***
## BsmtFinSF2         25.7586       8.8155   2.922              0.00358 **
## BsmtUnfSF          16.0895       5.5054   2.922              0.00357 **
## X1stFlrSF          49.7273       7.2829   6.828     0.00000000001695 ***
## X2ndFlrSF          51.2438       5.1996   9.855 < 0.0000000000000002 ***
## BedroomAbvGr   -12958.4912    2095.6401  -6.184     0.00000000099538 ***
## TotRmsAbvGrd     6806.0431    1510.9798   4.504     0.00000763977940 ***
## GarageCars       6445.6123    2192.9983   2.939              0.00338 **
## ScreenPorch        68.4666      21.6806   3.158              0.00165 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33620 on 806 degrees of freedom
##   (177 observations deleted due to missingness)
## Multiple R-squared:  0.8472, Adjusted R-squared:  0.8442
## F-statistic: 279.4 on 16 and 806 DF,  p-value: < 0.00000000000000022
```

```r
qqnorm(model3$residuals,col='green')
qqline(model3$residuals) ## plotting the residual values
```



Residual Plot vs Residual Line Chart

Variance

```
anova(model3) ## we are getting an accuracy of 84.72 from the model
```

```
## Analysis of Variance Table
##
## Response: SalePrice
##              Df        Sum Sq       Mean Sq   F value
## MSSubClass     1   60104013400   60104013400   53.1882
## LotFrontage    1  781913968163  781913968163  691.9436
## LotArea        1  209676962873  209676962873  185.5506
## OverallQual    1 3139931834321 3139931834321 2778.6378
## YearBuilt      1   59989935432   59989935432   53.0872
## YearRemodAdd   1   37180930942   37180930942   32.9027
## MasVnrArea     1  172783345318  172783345318  152.9022
## BsmtFinSF1     1  140992116346  140992116346  124.7690
## BsmtFinSF2     1    2021019950    2021019950    1.7885
## BsmtUnfSF      1   45672758737   45672758737   40.4175
## X1stFlrSF      1   38745158069   38745158069   34.2870
## X2ndFlrSF      1  287913767281  287913767281  254.7852
## BedroomAbvGr   1   28690874742   28690874742   25.3896
## TotRmsAbvGrd   1   24281412894   24281412894   21.4875
## GarageCars     1    9700562621    9700562621    8.5844
## ScreenPorch    1   11269497451   11269497451    9.9728
## Residuals    806  910800636548    1130025604
##                            Pr(>F)
## MSSubClass      0.0000000000007251 ***
## LotFrontage   < 0.00000000000000022 ***
## LotArea       < 0.00000000000000022 ***
## OverallQual   < 0.00000000000000022 ***
## YearBuilt       0.0000000000007609 ***
## YearRemodAdd    0.0000000137033394 ***
## MasVnrArea    < 0.00000000000000022 ***
## BsmtFinSF1    < 0.00000000000000022 ***
## BsmtFinSF2               0.181490
## BsmtUnfSF       0.0000000003431166 ***
## X1stFlrSF       0.0000000069206281 ***
## X2ndFlrSF     < 0.00000000000000022 ***
## BedroomAbvGr    0.0000005785525169 ***
## TotRmsAbvGrd    0.0000041529236424 ***
## GarageCars               0.003486 **
## ScreenPorch              0.001648 **
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

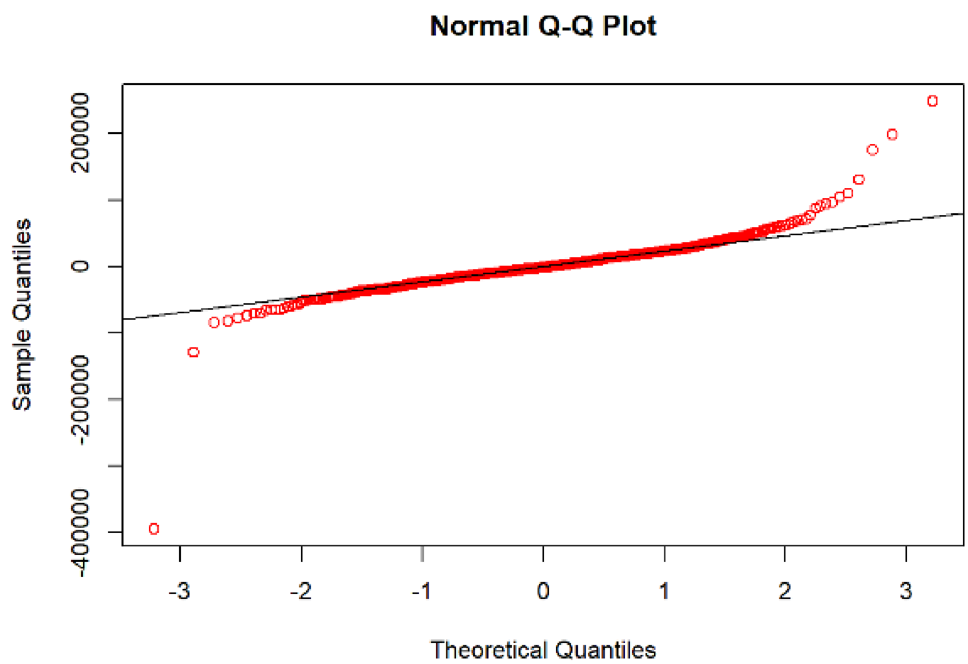
Model 4:

With 19 variables we predicted 84.72% accuracy.

```
## By considering the variables which are statistically significant by  taking the pr value cutoff as less than 0.1
model4_data<-train_1_[,c(2,3,4,5,7,8,9,10,11,12,13,14,15,20,22,24,25,26,31,36)]
model4<-lm(SalePrice~.,data = model4_data)
summary(model4) ## we are taking 19 variables
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = model4_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -394571  -15652   -1373   15382  247851
##
## Coefficients:
##                  Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)  -1174390.0162  160570.2278  -7.314    0.000000000000667 ***
## MSSubClass       -174.9992      35.3410  -4.952    0.000000909005208 ***
## LotFrontage        13.7313      67.3019   0.204              0.83839
## LotArea             0.6736       0.1483   4.543    0.000006472316248 ***
## OverallQual     19154.0214    1567.6363  12.218 < 0.0000000000000002 ***
## YearBuilt         206.6109      77.8776   2.653              0.00815 **
## YearRemodAdd      372.1942      83.6178   4.451    0.000009837036656 ***
## MasVnrArea         32.1311       7.2798   4.414    0.000011648536862 ***
## BsmtFinSF1         38.2922       5.9033   6.487    0.000000000159225 ***
## BsmtFinSF2         23.8680       9.0019   2.651              0.00818 **
## BsmtUnfSF          14.1629       5.7630   2.458              0.01421 *
## X1stFlrSF          45.0492       7.5923   5.934    0.000000004524380 ***
## X2ndFlrSF          49.0024       5.4203   9.041 < 0.0000000000000002 ***
## LowQualFinSF       51.0273      27.4906   1.856              0.06382 .
## BedroomAbvGr   -12307.3851    2308.5581  -5.331    0.000000129172647 ***
## TotRmsAbvGrd     6798.4697    1581.8457   4.298    0.000019520218205 ***
## GarageYrBlt       -19.6625      97.7687  -0.201              0.84067
## GarageCars       3666.0113    3763.3832   0.974              0.33031
## GarageArea         36.4896      13.1110   2.783              0.00552 **
## ScreenPorch        69.3546      21.7891   3.183              0.00152 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33630 on 751 degrees of freedom
##   (229 observations deleted due to missingness)
## Multiple R-squared:  0.8472, Adjusted R-squared:  0.8433
## F-statistic: 219.1 on 19 and 751 DF,  p-value: < 0.00000000000000022
```

```
qqnorm(model4$residuals,col='red')
qqline(model4$residuals) ## plotting the residual values
```



Residual Plotting

By comparison of al the models we suggest model3 as optimal model with 16 variables and accuracy of 84.72%.

Now, Let's apply the suggested model on the given test data.

Removing the NA's from test data set.

```
colMeans(is.na(test)) ## viewing na values of test data
```

```
##           Id    MSSubClass   LotFrontage       LotArea   OverallQual
##  0.000000000   0.000000000   0.186956522   0.000000000   0.000000000
##  OverallCond     YearBuilt  YearRemodAdd    MasVnrArea    BsmtFinSF1
##  0.000000000   0.000000000   0.000000000   0.004347826   0.000000000
##   BsmtFinSF2     BsmtUnfSF     X1stFlrSF     X2ndFlrSF  LowQualFinSF
##  0.000000000   0.000000000   0.000000000   0.000000000   0.000000000
## BsmtFullBath  BsmtHalfBath      FullBath      HalfBath  BedroomAbvGr
##  0.000000000   0.000000000   0.000000000   0.000000000   0.000000000
## KitchenAbvGr  TotRmsAbvGrd    Fireplaces   GarageYrBlt    GarageCars
##  0.000000000   0.000000000   0.000000000   0.054347826   0.000000000
##   GarageArea    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##  0.000000000   0.000000000   0.000000000   0.000000000   0.000000000
##  ScreenPorch      PoolArea       MiscVal        MoSold        YrSold
##  0.000000000   0.000000000   0.000000000   0.000000000   0.000000000
```

Predicting the price values using the best model

```
test_predict<-predict(model3,test1)
```

```
test1$predicted_value<-test_predict
```

Estimating the Prediction interval, the fitted value is the same as before, but the interval is wider. This is due to the additional term in the standard error of prediction

```
head(predict(model3,test1,interval = "prediction",level = 0.95))
```

```
##         fit        lwr      upr
## 1  50303.67 -16717.520 117324.9
## 2  64531.01  -1818.914 130880.9
## 3 246422.74 180039.543 312805.9
## 4 161246.39  94526.127 227966.7
## 5 216642.90 149769.590 283516.2
## 6 124048.71  57806.900 190290.5
```

Confidence Interval:

```
head(confint(model3,level = 0.9)) ## confidence levels
```

```
##                          5 %           95 %
## (Intercept) -1503442.4654005 -1027423.119021
## MSSubClass      -203.9427916      -98.062747
## LotFrontage      -58.2553863      154.990667
## LotArea            0.4389822        0.925397
## OverallQual    15355.8290387    20206.738987
## YearBuilt        135.4908942      313.682645
```

Conclusions

We have predicted house rates by building linear regressing using 16 variables and got accuracy of 84.72%.

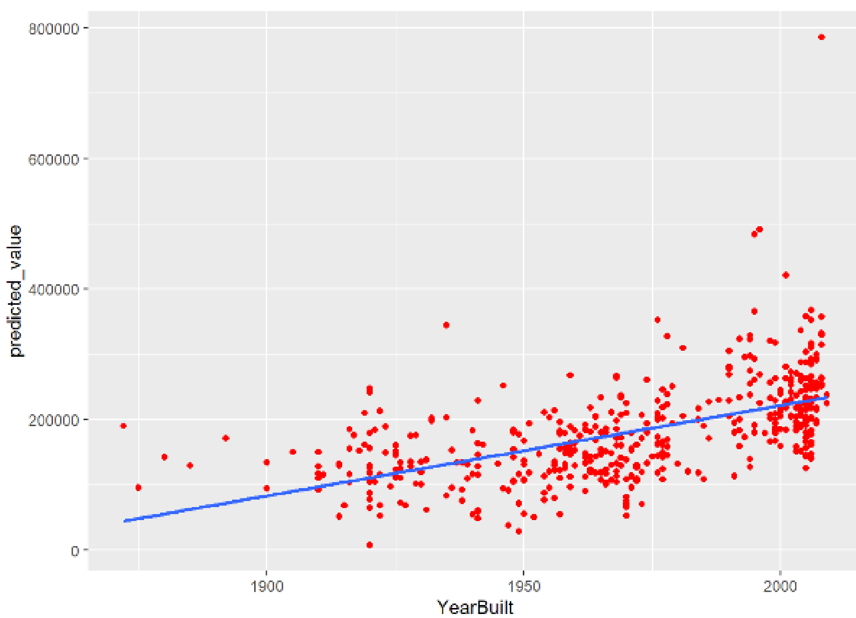
By above scenarios we can say the following:

The prediction value highly varies by

- Built Year
- Feet of street connected to property
- Overall material and finish of the house

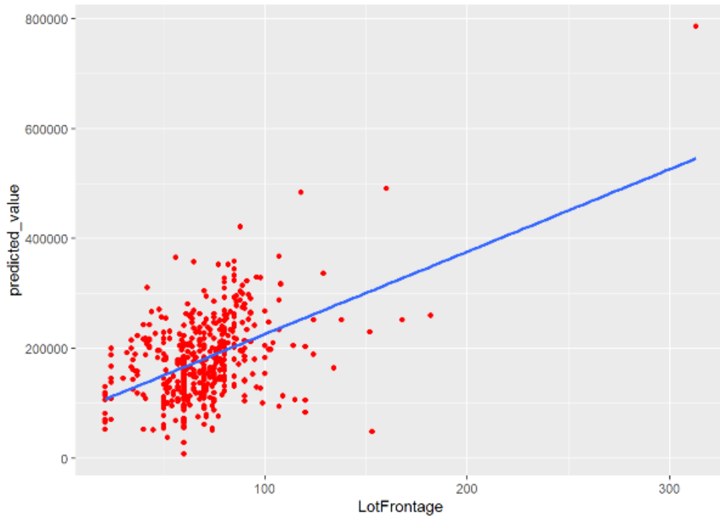
Plotting's:

```
## plotting between prdicted value and Year built
ggplot(data = test1, aes(x = YearBuilt, y = predicted_value)) +
  geom_point(color='red') +
  geom_smooth(method = "lm", se = FALSE)
```

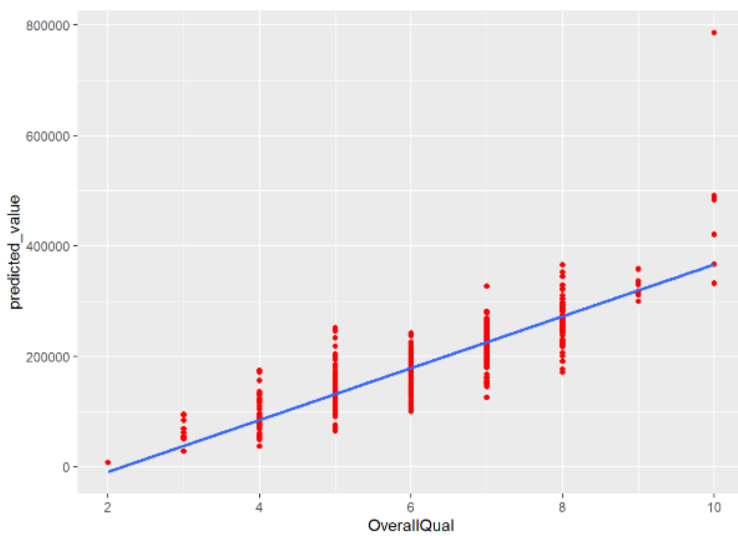


By the above graph we can say that less the age of home the better price.

```
## plotting between predicted values and Lot Frontage
ggplot(data = test1, aes(x = LotFrontage, y = predicted_value)) +
  geom_point(color='red') +
  geom_smooth(method = "lm", se = FALSE)
```




```
## plotting between predicted values and ovrall quality
ggplot(data = test1, aes(x = OverallQual, y = predicted_value)) +
  geom_point(color='red') +
  geom_smooth(method = "lm", se = FALSE)
```



Better the Quality and finishing better the pricing.

Insights

- Limit the number of parameters and start building the model.
- Understand the length of time it takes to fit a model before running it.
- Picking the best models with minimum number of variables.

****