House Price Prediction

Sakshi

2023-12-09

Loading the required packages

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(rattle)
```

```
Loading required package: bitops
Rattle: A free graphical interface for data science with R.
Version 5.5.1 Copyright (c) 2006-2021 Togaware Pty Ltd.
Type 'rattle()' to shake, rattle, and roll your data.
```

```
library(caret)
```

```
Loading required package: lattice

Attaching package: 'caret'

The following object is masked from 'package:purrr':

    lift
```

```
library(class)
library(rpart)
library(corrplot)
```

```
corrplot 0.92 loaded
```

```r
library(pander)
```

Reading both CSV and Excel file

```r
df.predict <- readxl::read_excel("/Users/sakshibansal/Downloads/BA-Predict-2.xlsx")
df.houseprice <- read.csv("/Users/sakshibansal/Downloads/House_Prices.csv")
```

Descriptive analytics

```r
# Checking if data is imported correctly
head(df.predict)
```

```
# A tibble: 6 x 13
  LotArea OverallQual YearBuilt YearRemodAdd BsmtFinSF1 FullBath HalfBath
    <dbl>       <dbl>     <dbl>        <dbl>      <dbl>    <dbl>    <dbl>
1    7340           4      1971         1971        322        1        0
2    8712           5      1957         2000        860        1        0
3    7875           7      2003         2003          0        2        1
4   14859           7      2006         2006          0        2        0
5    6173           5      1967         1967        599        1        0
6    9920           5      1954         1954        354        1        0
# i 6 more variables: BedroomAbvGr <dbl>, TotRmsAbvGrd <dbl>, Fireplaces <dbl>,
#   GarageArea <dbl>, YrSold <dbl>, SalePrice <dbl>
```

```r
head(df.houseprice)
```

```
  LotArea OverallQual YearBuilt YearRemodAdd BsmtFinSF1 FullBath HalfBath
1    8450           7      2003         2003        706        2        1
2    9600           6      1976         1976        978        2        0
3   11250           7      2001         2002        486        2        1
4    9550           7      1915         1970        216        1        0
5   14260           8      2000         2000        655        2        1
6   14115           5      1993         1995        732        1        1
  BedroomAbvGr TotRmsAbvGrd Fireplaces GarageArea YrSold SalePrice
1            3            8          0        548   2008    208500
2            3            6          1        460   2007    181500
3            3            6          1        608   2008    223500
4            3            7          1        642   2006    140000
5            4            9          1        836   2008    250000
6            1            5          0        480   2009    143000
```

```r
# Checking the dimensions of the data
dim(df.predict)
```

```
[1] 90 13
```

```r
dim(df.houseprice)
```

```
[1] 900  13
```

```r
# Checking the structure of the data
str(df.predict)
```

```
tibble [90 x 13] (S3: tbl_df/tbl/data.frame)
 $ LotArea     : num [1:90] 7340 8712 7875 14859 6173 ...
 $ OverallQual : num [1:90] 4 5 7 7 5 5 8 7 5 6 ...
 $ YearBuilt   : num [1:90] 1971 1957 2003 2006 1967 ...
 $ YearRemodAdd: num [1:90] 1971 2000 2003 2006 1967 ...
 $ BsmtFinSF1  : num [1:90] 322 860 0 0 599 354 63 223 301 0 ...
 $ FullBath    : num [1:90] 1 1 2 2 1 1 2 1 1 2 ...
 $ HalfBath    : num [1:90] 0 0 1 0 0 0 0 1 0 1 ...
 $ BedroomAbvGr: num [1:90] 2 2 3 3 3 3 3 3 2 3 ...
 $ TotRmsAbvGrd: num [1:90] 4 5 8 7 6 6 8 6 5 8 ...
 $ Fireplaces  : num [1:90] 0 0 1 1 0 0 1 1 0 1 ...
 $ GarageArea  : num [1:90] 684 756 393 690 288 280 865 180 484 390 ...
 $ YrSold      : num [1:90] 2007 2009 2006 2006 2007 ...
 $ SalePrice   : num [1:90] 110000 153000 180000 240000 125500 ...
```

```r
str(df.houseprice)
```

```
'data.frame':   900 obs. of  13 variables:
 $ LotArea     : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
 $ OverallQual : int  7 6 7 7 8 5 8 7 7 5 ...
 $ YearBuilt   : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
 $ YearRemodAdd: int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
 $ BsmtFinSF1  : int  706 978 486 216 655 732 1369 859 0 851 ...
 $ FullBath    : int  2 2 2 1 2 1 2 2 2 1 ...
 $ HalfBath    : int  1 0 1 0 1 1 0 1 0 0 ...
 $ BedroomAbvGr: int  3 3 3 3 4 1 3 3 2 2 ...
 $ TotRmsAbvGrd: int  8 6 6 7 9 5 7 7 8 5 ...
 $ Fireplaces  : int  0 1 1 1 1 0 1 2 2 2 ...
 $ GarageArea  : int  548 460 608 642 836 480 636 484 468 205 ...
 $ YrSold      : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
 $ SalePrice   : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

```r
# Understanding the data by looking at the summary
summary(df.predict)
```

```
    LotArea       OverallQual   YearBuilt     YearRemodAdd    BsmtFinSF1    
 Min.   : 1300   Min.   :2    Min.   :1890   Min.   :1950   Min.   :   0.0  
 1st Qu.: 7493   1st Qu.:5    1st Qu.:1958   1st Qu.:1966   1st Qu.:   0.0  
 Median : 9380   Median :6    Median :1976   Median :1994   Median : 407.5  
 Mean   : 9713   Mean   :6    Mean   :1974   Mean   :1985   Mean   : 426.1  
 3rd Qu.:11629   3rd Qu.:7    3rd Qu.:2002   3rd Qu.:2004   3rd Qu.: 687.0  
 Max.   :27650   Max.   :9    Max.   :2009   Max.   :2010   Max.   :1646.0  
    FullBath        HalfBath       BedroomAbvGr    TotRmsAbvGrd   
 Min.   :0.000   Min.   :0.0000   Min.   :1.000   Min.   : 4.000  
 1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.250   1st Qu.: 5.250  
 Median :2.000   Median :0.0000   Median :3.000   Median : 6.000  
 Mean   :1.578   Mean   :0.3778   Mean   :2.967   Mean   : 6.633  
 3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.: 8.000  
 Max.   :2.000   Max.   :2.0000   Max.   :5.000   Max.   :12.000  
```

```
  Fireplaces       GarageArea        YrSold        SalePrice
 Min.   :0.0000   Min.   :  0.0   Min.   :2006   Min.   : 35311
 1st Qu.:0.0000   1st Qu.:388.5   1st Qu.:2007   1st Qu.:132475
 Median :0.0000   Median :491.0   Median :2008   Median :166250
 Mean   :0.4333   Mean   :475.4   Mean   :2008   Mean   :172587
 3rd Qu.:1.0000   3rd Qu.:604.8   3rd Qu.:2009   3rd Qu.:200725
 Max.   :2.0000   Max.   :871.0   Max.   :2010   Max.   :395192
```

```r
summary(df.houseprice)
```

```
    LotArea        OverallQual       YearBuilt     YearRemodAdd
 Min.   :  1491   Min.   : 1.000   Min.   :1880   Min.   :1950
 1st Qu.:  7585   1st Qu.: 5.000   1st Qu.:1954   1st Qu.:1968
 Median :  9442   Median : 6.000   Median :1973   Median :1994
 Mean   : 10795   Mean   : 6.136   Mean   :1971   Mean   :1985
 3rd Qu.: 11618   3rd Qu.: 7.000   3rd Qu.:2000   3rd Qu.:2004
 Max.   :215245   Max.   :10.000   Max.   :2010   Max.   :2010
  BsmtFinSF1        FullBath        HalfBath        BedroomAbvGr
 Min.   :   0.0   Min.   :0.000   Min.   :0.0000   Min.   :0.000
 1st Qu.:   0.0   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000
 Median : 384.0   Median :2.000   Median :0.0000   Median :3.000
 Mean   : 446.5   Mean   :1.564   Mean   :0.3856   Mean   :2.843
 3rd Qu.: 728.8   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000
 Max.   :2260.0   Max.   :3.000   Max.   :2.0000   Max.   :8.000
 TotRmsAbvGrd      Fireplaces       GarageArea         YrSold
 Min.   : 2.000   Min.   :0.0000   Min.   :   0.0   Min.   :2006
 1st Qu.: 5.000   1st Qu.:0.0000   1st Qu.: 336.0   1st Qu.:2007
 Median : 6.000   Median :1.0000   Median : 480.0   Median :2008
 Mean   : 6.482   Mean   :0.6278   Mean   : 472.6   Mean   :2008
 3rd Qu.: 7.000   3rd Qu.:1.0000   3rd Qu.: 576.0   3rd Qu.:2009
 Max.   :14.000   Max.   :3.0000   Max.   :1390.0   Max.   :2010
   SalePrice
 Min.   : 34900
 1st Qu.:130000
 Median :163000
 Mean   :183108
 3rd Qu.:216878
 Max.   :755000
```

DATA PREPRATION:

Checking for missing values in the dataset:

```r
#Checking for training set
colSums(is.na(df.houseprice))
```

```
     LotArea  OverallQual    YearBuilt YearRemodAdd   BsmtFinSF1     FullBath
           0            0            0            0            0            0
    HalfBath BedroomAbvGr TotRmsAbvGrd   Fireplaces   GarageArea       YrSold
           0            0            0            0            0            0
   SalePrice
           0
```

```r
#Checking for testing set
colSums(is.na(df.predict))
```

```
     LotArea  OverallQual    YearBuilt YearRemodAdd   BsmtFinSF1     FullBath 
           0            0            0            0            0            0 
    HalfBath BedroomAbvGr TotRmsAbvGrd   Fireplaces   GarageArea       YrSold 
           0            0            0            0            0            0 
   SalePrice 
           0 
```

Hence, there are no missing values in our data

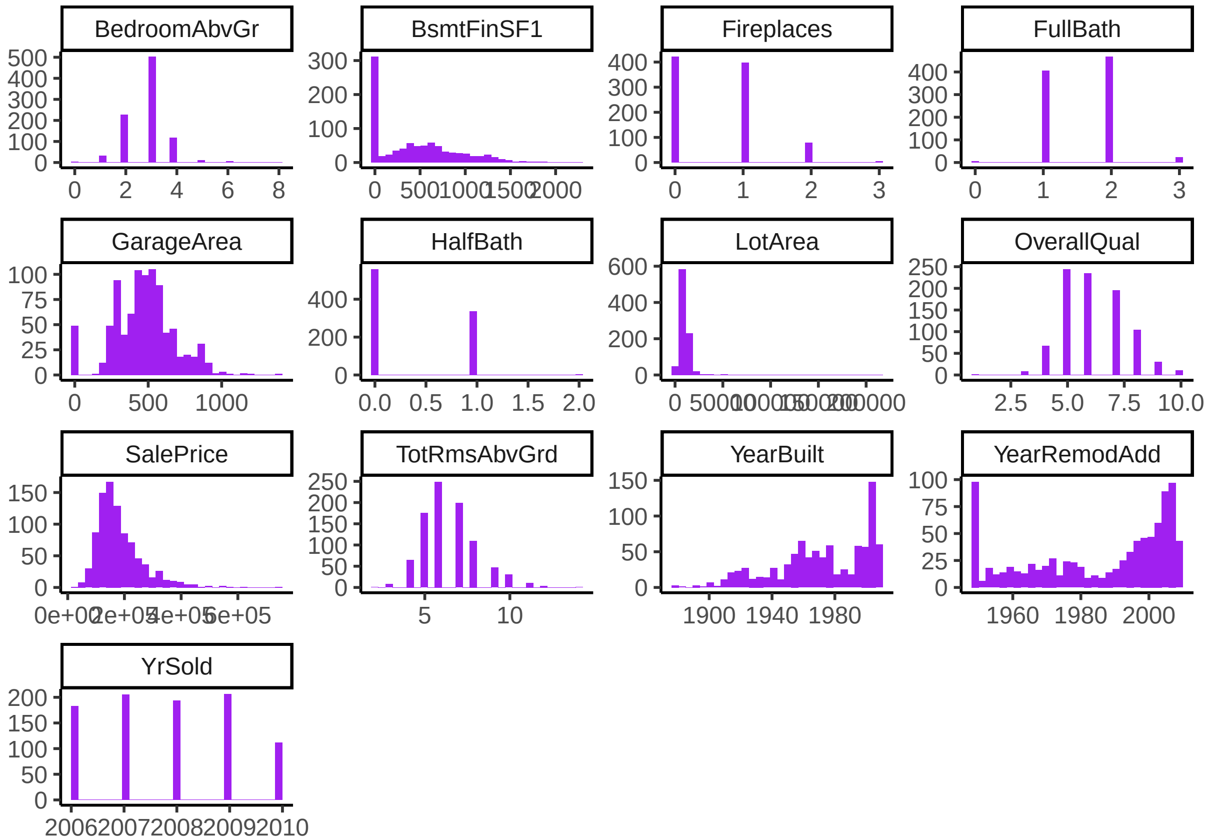
Some of the categorical variables in the data are of type 'integer' and should be as factors instead to run the models the right way.

DATA EXPLORATION

Histogram plots of all variables:

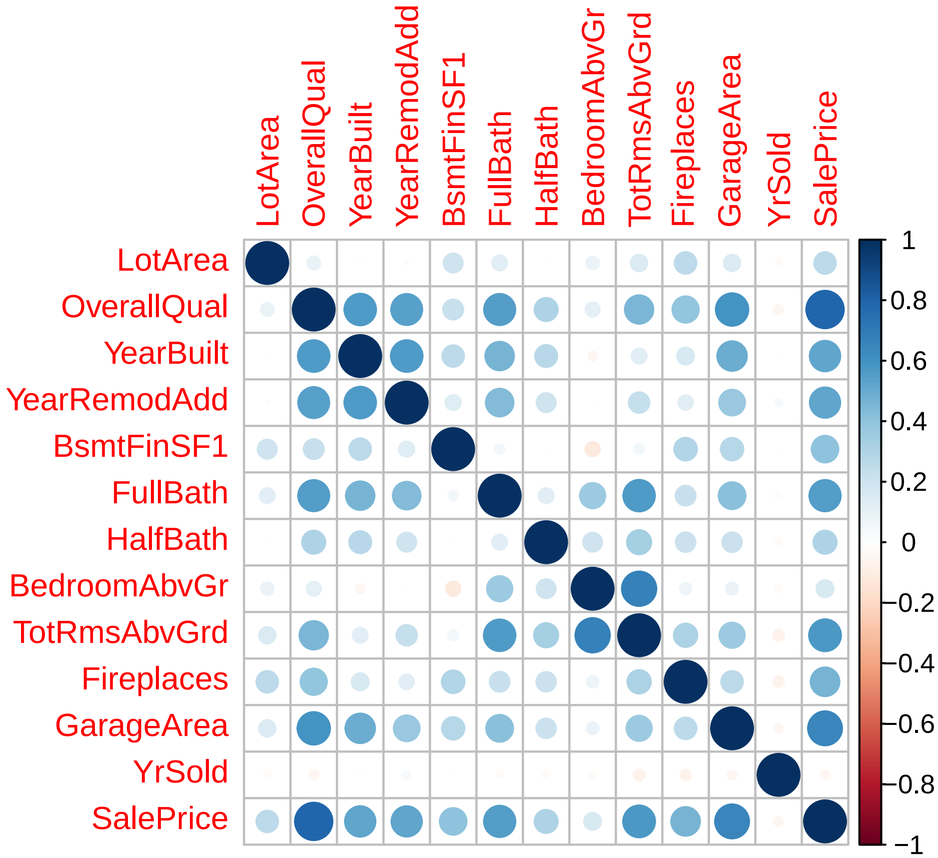
```r
Hist_data <- df.houseprice %>%
  gather(key = "Variable", value = "Value")

ggplot(Hist_data, aes(x = Value)) +
  geom_histogram(fill = "purple", bins = 30) +
  facet_wrap(~ Variable, scales = 'free') +
  theme_classic() +
  theme(aspect.ratio = 0.5, axis.title = element_blank(), panel.grid = element_blank())
```



Trying corrplot to make sure no variables are correlated:

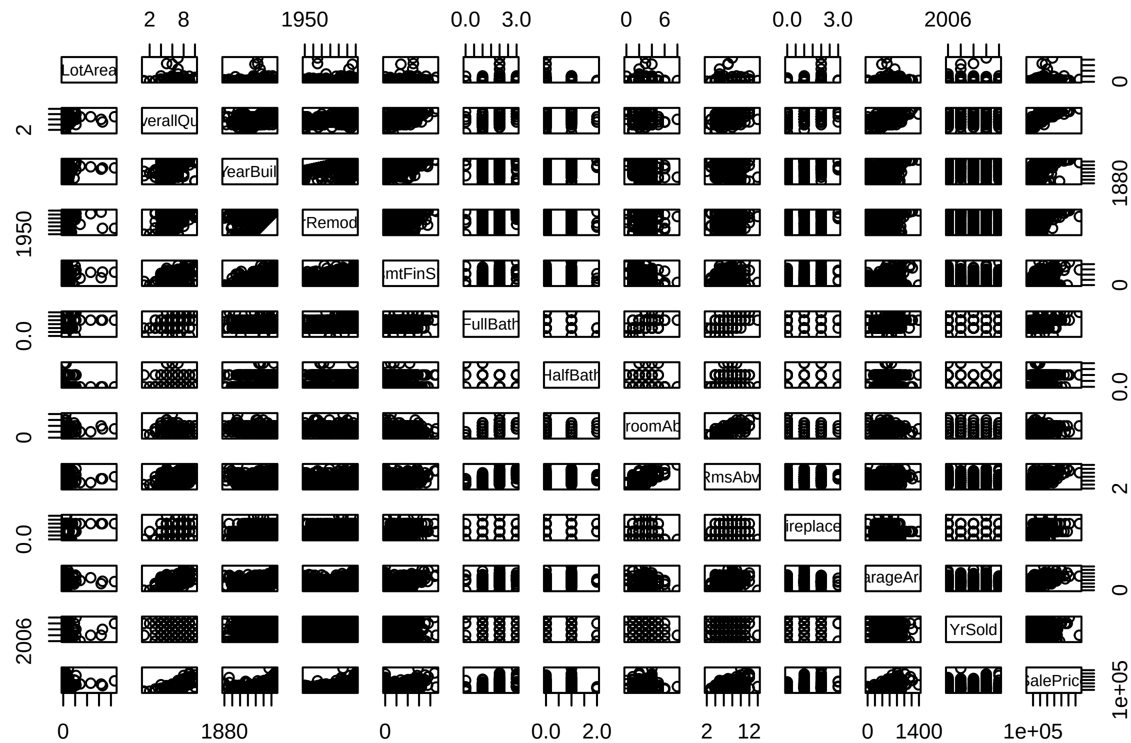
```
corrplot(cor(df.houseprice))
```



As,we can see above in the corrplot, the TotRmsAbvGrd is highly correlated to BedroomAbvGrd and Full bath, which has created a situation of **Multicollinearity**.Due to this, the coefficient estimates of the regression model can be unstable and highly sensitive to changes in the model leading to incorrect conclusions about the significance of the predictors. Hence, we decided to remove TotRmsAbvGrd from our model to predict correctly.

Making pairs to have a better understanding of variables:

```
pairs(df.houseprice)
```



Converting categorical variables as factors-

```
df.houseprice[,c("OverallQual","FullBath","HalfBath","BedroomAbvGr","TotRmsAbvGrd","Fireplaces","YrSold
```

Converting categorical variables as factors in our test data as well-

```
df.predict[,c("OverallQual","FullBath","HalfBath","BedroomAbvGr","TotRmsAbvGrd","Fireplaces","YrSold")]
```

Regression Analytics:

```
l=lm(SalePrice~.,data = df.houseprice)
summary(l)
```

```
Call:
lm(formula = SalePrice ~ ., data = df.houseprice)

Residuals:
    Min      1Q  Median      3Q     Max
-297845  -15887       0   13331  253342
```

```
Coefficients: (1 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.234e+06  1.525e+05  -8.092 2.00e-15 ***
LotArea         6.794e-01  9.643e-02   7.046 3.79e-12 ***
OverallQual2   -8.571e+03  5.049e+04  -0.170 0.865245
OverallQual3    5.566e+03  4.157e+04   0.134 0.893520
OverallQual4    1.026e+04  4.027e+04   0.255 0.798941
OverallQual5    1.634e+04  4.023e+04   0.406 0.684736
OverallQual6    2.656e+04  4.028e+04   0.659 0.509819
OverallQual7    5.068e+04  4.036e+04   1.256 0.209605
OverallQual8    8.967e+04  4.068e+04   2.204 0.027785 *
OverallQual9    1.543e+05  4.093e+04   3.770 0.000174 ***
OverallQual10   1.731e+05  4.201e+04   4.120 4.16e-05 ***
YearBuilt       2.042e+02  5.545e+01   3.683 0.000245 ***
YearRemodAdd    4.385e+02  6.972e+01   6.290 5.07e-10 ***
BsmtFinSF1      2.830e+01  2.786e+00  10.160  < 2e-16 ***
FullBath1       1.264e+04  2.341e+04   0.540 0.589207
FullBath2       1.657e+04  2.365e+04   0.701 0.483771
FullBath3       4.500e+04  2.504e+04   1.797 0.072661 .
HalfBath1       6.966e+03  2.612e+03   2.667 0.007798 **
HalfBath2      -1.205e+04  1.765e+04  -0.683 0.494855
BedroomAbvGr1  -2.293e+02  2.515e+04  -0.009 0.992727
BedroomAbvGr2  -1.264e+04  2.440e+04  -0.518 0.604662
BedroomAbvGr3  -1.722e+04  2.461e+04  -0.699 0.484476
BedroomAbvGr4  -9.604e+03  2.503e+04  -0.384 0.701232
BedroomAbvGr5  -3.198e+04  2.794e+04  -1.145 0.252710
BedroomAbvGr6  -8.017e+04  2.935e+04  -2.732 0.006429 **
BedroomAbvGr8   1.055e+05  5.877e+04   1.796 0.072908 .
TotRmsAbvGrd3   7.918e+02  5.217e+04   0.015 0.987893
TotRmsAbvGrd4   2.404e+04  5.051e+04   0.476 0.634173
TotRmsAbvGrd5   3.209e+04  5.070e+04   0.633 0.527010
TotRmsAbvGrd6   3.928e+04  5.082e+04   0.773 0.439709
TotRmsAbvGrd7   4.977e+04  5.090e+04   0.978 0.328411
TotRmsAbvGrd8   5.755e+04  5.100e+04   1.128 0.259483
TotRmsAbvGrd9   7.134e+04  5.123e+04   1.393 0.164083
TotRmsAbvGrd10  1.101e+05  5.157e+04   2.135 0.033048 *
TotRmsAbvGrd11  6.465e+04  5.223e+04   1.238 0.216145
TotRmsAbvGrd12  1.927e+05  5.407e+04   3.563 0.000386 ***
TotRmsAbvGrd14         NA         NA      NA       NA
Fireplaces1     1.159e+04  2.588e+03   4.479 8.52e-06 ***
Fireplaces2     2.404e+04  4.527e+03   5.309 1.40e-07 ***
Fireplaces3     3.133e+04  1.620e+04   1.934 0.053467 .
GarageArea      5.199e+01  6.960e+00   7.469 1.99e-13 ***
YrSold2007      5.229e+03  3.268e+03   1.600 0.109939
YrSold2008      2.013e+03  3.313e+03   0.608 0.543591
YrSold2009      1.497e+02  3.276e+03   0.046 0.963561
YrSold2010      4.321e+03  3.860e+03   1.119 0.263264
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31520 on 856 degrees of freedom
Multiple R-squared:  0.859, Adjusted R-squared:  0.8519
F-statistic: 121.3 on 43 and 856 DF,  p-value: < 2.2e-16
```

We can get the amount of variability explained by each variable by anova analysis:

```r
anova(l)
```

```
Analysis of Variance Table

Response: SalePrice
              Df     Sum Sq    Mean Sq  F value    Pr(>F)    
LotArea        1 4.2155e+11 4.2155e+11 424.2351 < 2.2e-16 ***
OverallQual    9 3.9715e+12 4.4127e+11 444.0856 < 2.2e-16 ***
YearBuilt      1 7.8814e+10 7.8814e+10  79.3167 < 2.2e-16 ***
YearRemodAdd   1 4.1278e+10 4.1278e+10  41.5407 1.926e-10 ***
BsmtFinSF1     1 1.4466e+11 1.4466e+11 145.5828 < 2.2e-16 ***
FullBath       3 1.1783e+11 3.9275e+10  39.5255 < 2.2e-16 ***
HalfBath       2 7.4940e+10 3.7470e+10  37.7088 < 2.2e-16 ***
BedroomAbvGr   7 5.6007e+10 8.0009e+09   8.0519 1.676e-09 ***
TotRmsAbvGrd  10 1.8087e+11 1.8087e+10  18.2020 < 2.2e-16 ***
Fireplaces     3 3.4832e+10 1.1611e+10  11.6846 1.651e-07 ***
GarageArea     1 5.4552e+10 5.4552e+10  54.8993 3.044e-13 ***
YrSold         4 3.9956e+09 9.9891e+08   1.0053    0.4038    
Residuals    856 8.5058e+11 9.9367e+08                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above summary and anova analysis, we conclude that FullBath and YrSold variables are not significant.Also, we remove TotRmsAbvGrd due to multicollinearity. Hence, re-running the model on significant variables only.

```r
l=lm(SalePrice~.,data = df.houseprice[,-c(6,9,12)])
summary(l)
```

```
Call:
lm(formula = SalePrice ~ ., data = df.houseprice[, -c(6, 9, 12)])

Residuals:
    Min      1Q  Median      3Q     Max 
-300583  -17524   -1175   15381  270902 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.280e+06  1.515e+05  -8.446  < 2e-16 ***
LotArea        7.549e-01  1.040e-01   7.258 8.70e-13 ***
OverallQual2   6.333e+03  4.217e+04   0.150 0.880662    
OverallQual3   2.005e+04  2.774e+04   0.723 0.470119    
OverallQual4   1.763e+04  2.550e+04   0.691 0.489437    
OverallQual5   2.181e+04  2.548e+04   0.856 0.392126    
OverallQual6   3.376e+04  2.553e+04   1.322 0.186403    
OverallQual7   6.289e+04  2.580e+04   2.438 0.014975 *  
OverallQual8   1.083e+05  2.610e+04   4.151 3.64e-05 ***
OverallQual9   1.838e+05  2.680e+04   6.857 1.33e-11 ***
OverallQual10  2.300e+05  2.825e+04   8.141 1.35e-15 ***
YearBuilt      1.250e+02  5.546e+01   2.254 0.024420 *  
```

```
YearRemodAdd   5.422e+02  7.419e+01   7.309 6.09e-13 ***
BsmtFinSF1     2.708e+01  2.985e+00   9.073  < 2e-16 ***
HalfBath1      1.035e+04  2.692e+03   3.846 0.000129 ***
HalfBath2     -1.785e+04  1.578e+04  -1.132 0.258089
BedroomAbvGr1  1.513e+04  2.105e+04   0.719 0.472371
BedroomAbvGr2  1.599e+04  2.029e+04   0.788 0.430946
BedroomAbvGr3  2.122e+04  2.019e+04   1.051 0.293705
BedroomAbvGr4  5.065e+04  2.034e+04   2.490 0.012971 *
BedroomAbvGr5  4.927e+04  2.302e+04   2.140 0.032643 *
BedroomAbvGr6  4.194e+04  2.472e+04   1.697 0.090113 .
BedroomAbvGr8  1.097e+05  4.007e+04   2.739 0.006292 **
Fireplaces1    1.423e+04  2.770e+03   5.139 3.41e-07 ***
Fireplaces2    3.147e+04  4.839e+03   6.503 1.33e-10 ***
Fireplaces3    3.251e+04  1.746e+04   1.862 0.062986 .
GarageArea     6.570e+01  7.397e+00   8.881  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34320 on 873 degrees of freedom
Multiple R-squared:  0.8295,    Adjusted R-squared:  0.8244
F-statistic: 163.3 on 26 and 873 DF,  p-value: < 2.2e-16
```

Now, as we have the right model with only significant variables, we can use this method to make predictions on test data.

```r
lt=predict(l,df.predict)
#Summary of predictions made
summary(lt)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  56945  128440  167092  177741  208449  382960
```

```r
#Displaying head of predicted values for demonstration
head(lt)
```

```
       1        2        3        4        5        6
128266.9 166756.5 197259.8 213691.4 115609.3 102603.5
```

Calculating errors of predictions:

```r
#RMSE calculation
RMSE(lt,df.predict$SalePrice)
```

```
[1] 23991.65
```

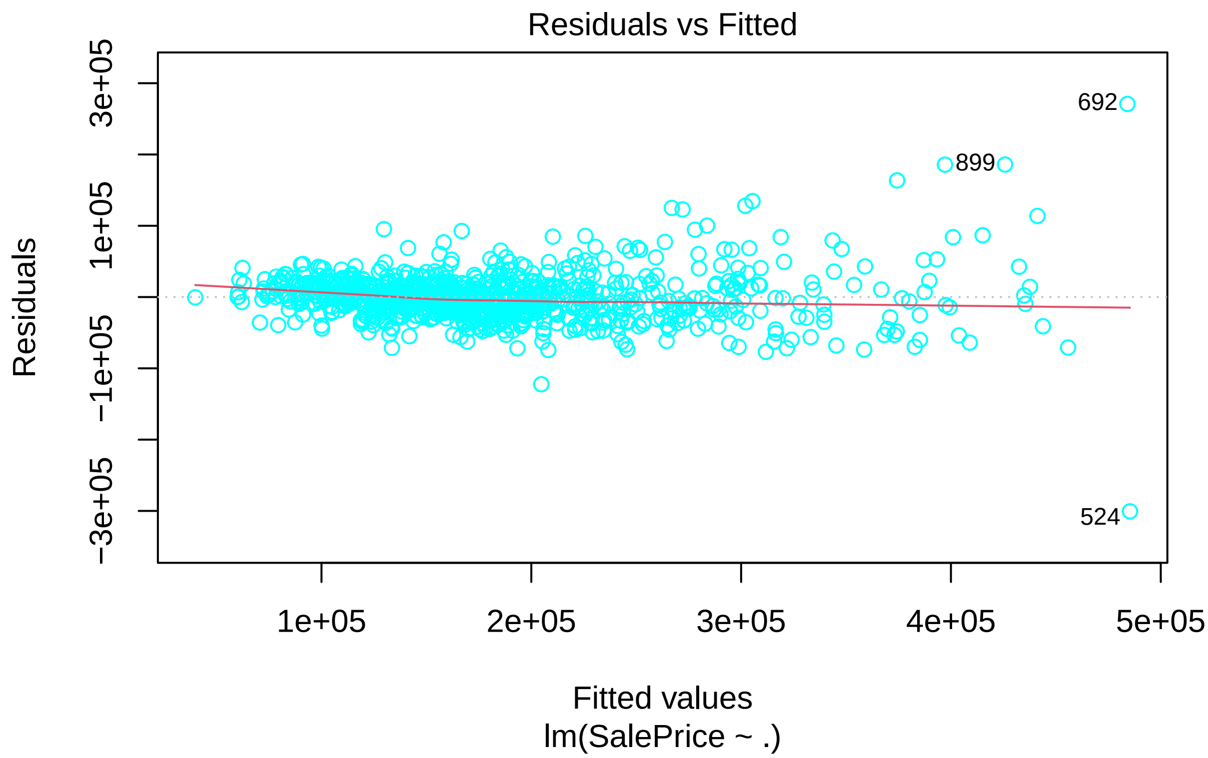
```r
#MAE calculation
MAE(lt,df.predict$SalePrice)
```

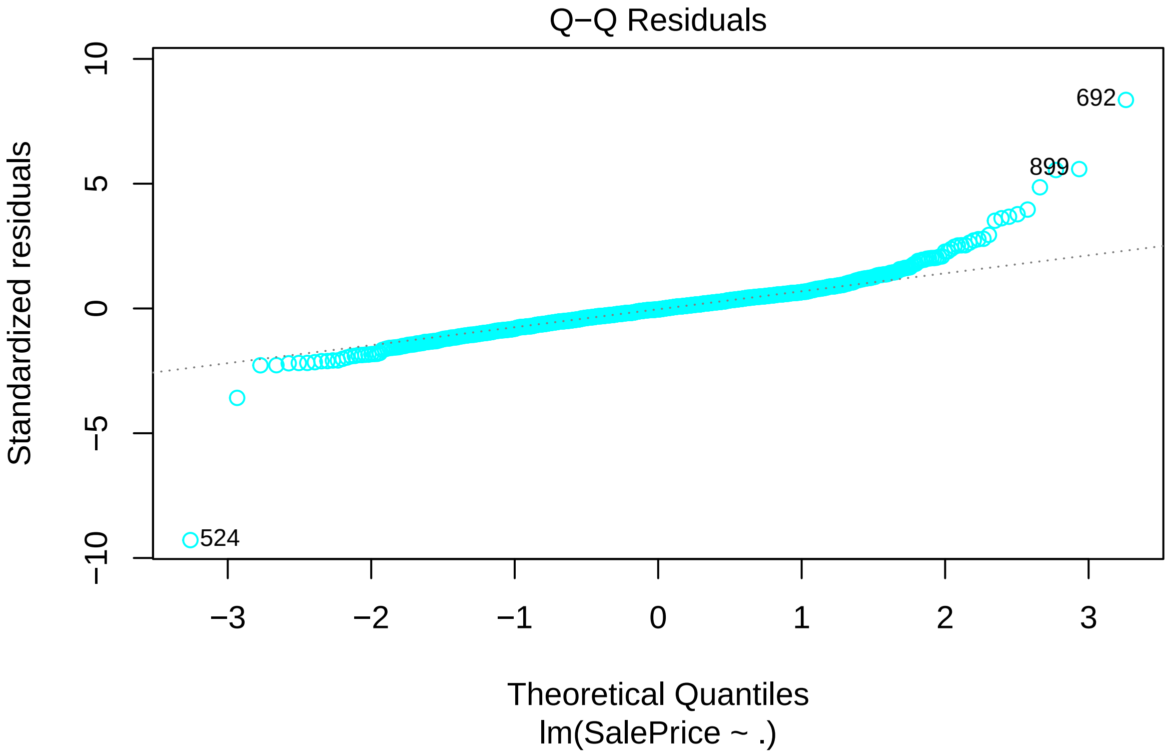
```
[1] 18945.79
```

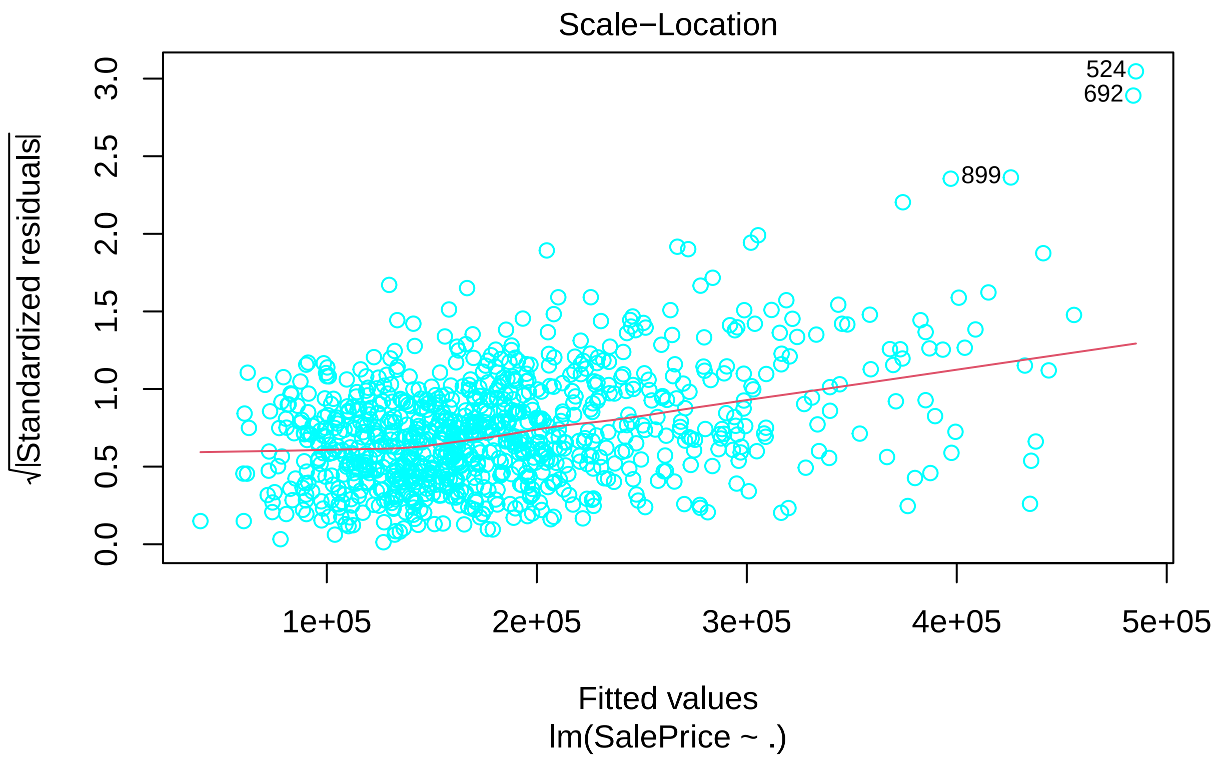
Plotting the model to check its efficiency:

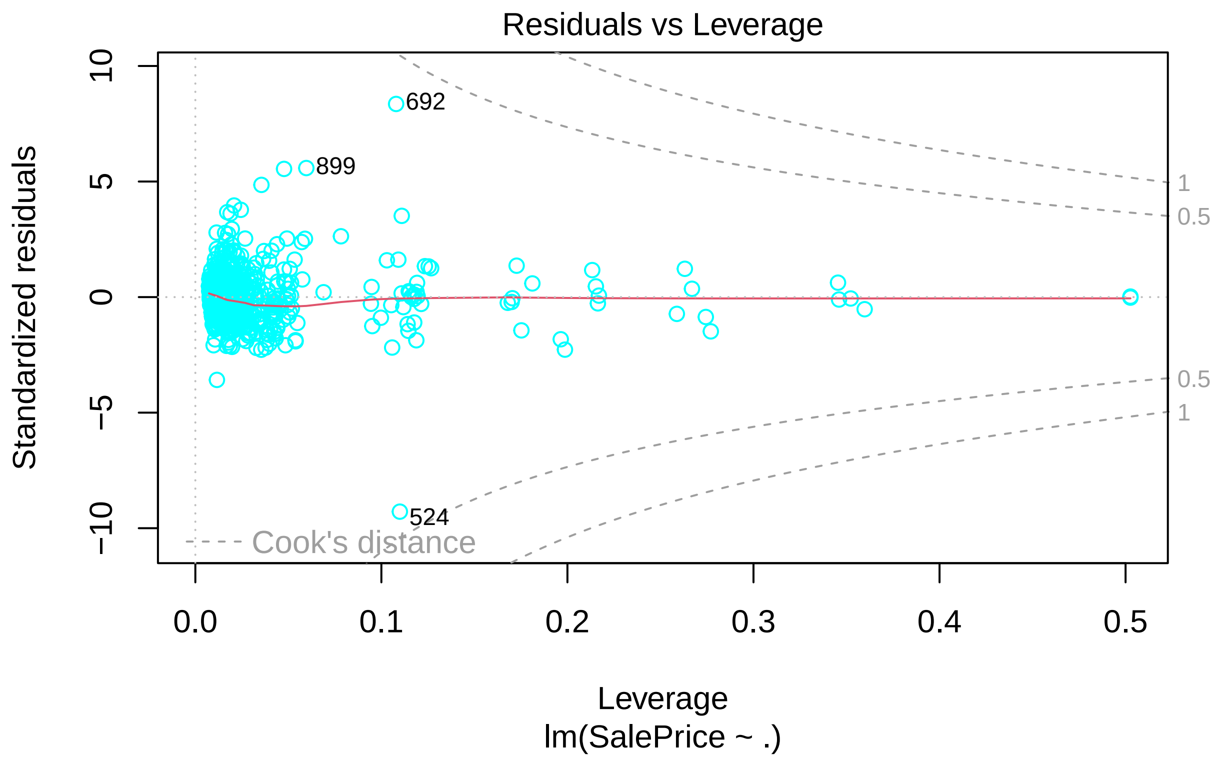
```r
plot(l,col='cyan')
```

```
Warning: not plotting observations with leverage one:
  636, 637
```




Q−Q Residuals
Standardized residuals
10
5
0
−5
−10
692
899
524
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(SalePrice ~ .)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
524
692
899
1e+05
2e+05
3e+05
4e+05
5e+05
Fitted values
lm(SalePrice ~ .)



The scatter plot reveals that the relationship between the fitted values and residuals is not entirely random; there appears to be some pattern, indicating potential issues with the model. Additionally, the Q-Q plot shows a S curve with outliers on both sides which means that our data is skewed resulting in deviations from the expected straight line as proved above, suggesting that the residuals might not follow a normal distribution. These observations indicate that the linear regression model may not fully meet the assumptions. As a result, we would further explore another model called decision tree for the same dataset.

Decision Tree:

```
d=rpart(SalePrice~.,data=df.houseprice,method = 'anova',control = rpart.control(minsplit = 30))
summary(d)
```

```
Call:
rpart(formula = SalePrice ~ ., data = df.houseprice, method = "anova",
    control = rpart.control(minsplit = 30))
  n= 900

           CP nsplit rel error    xerror       xstd
1  0.47785266      0 1.0000000 1.0023280 0.08999933
2  0.11551089      1 0.5221473 0.5245099 0.04645066
3  0.05814644      2 0.4066365 0.4094318 0.04398804
```

```
4  0.02917450      3 0.3484900 0.3876460 0.03465391
5  0.01958114      4 0.3193155 0.3559891 0.03375176
6  0.01801227      5 0.2997344 0.3552405 0.03311498
7  0.01429023      6 0.2817221 0.3515482 0.03287674
8  0.01200980      7 0.2674319 0.3408864 0.02984827
9  0.01183168      8 0.2554221 0.3380188 0.02982936
10 0.01000000      9 0.2435904 0.3294361 0.02974666

Variable importance
 OverallQual   GarageArea    YearBuilt   BsmtFinSF1 TotRmsAbvGrd YearRemodAdd
          52           13           10            9            5            4
    FullBath BedroomAbvGr      LotArea   Fireplaces
           3            1            1            1

Node number 1: 900 observations,    complexity param=0.4778527
  mean=183107.9, MSE=6.701496e+09
  left son=2 (754 obs) right son=3 (146 obs)
  Primary splits:
      OverallQual  splits as  LLLLLLLRRR,  improve=0.4778527, (0 missing)
      YearBuilt    < 1984.5  to the left,  improve=0.3410164, (0 missing)
      GarageArea   < 675.5   to the left,  improve=0.3352460, (0 missing)
      FullBath     splits as  RLRR,        improve=0.2777250, (0 missing)
      YearRemodAdd < 1983.5  to the left,  improve=0.2410551, (0 missing)
  Surrogate splits:
      GarageArea   < 679     to the left,   agree=0.891, adj=0.329, (0 split)
      YearBuilt    < 2005.5  to the left,   agree=0.863, adj=0.158, (0 split)
      BsmtFinSF1   < 1336    to the left,   agree=0.860, adj=0.137, (0 split)
      YearRemodAdd < 2007.5  to the left,   agree=0.850, adj=0.075, (0 split)
      TotRmsAbvGrd splits as  LLLLLLLLLRRL, agree=0.846, adj=0.048, (0 split)

Node number 2: 754 observations,    complexity param=0.1155109
  mean=158206.5, MSE=2.548301e+09
  left son=4 (558 obs) right son=5 (196 obs)
  Primary splits:
      OverallQual  splits as  LLLLLLR---,  improve=0.3625894, (0 missing)
      FullBath     splits as  LLRR,        improve=0.3232482, (0 missing)
      YearBuilt    < 1984.5  to the left,  improve=0.2933600, (0 missing)
      GarageArea   < 387     to the left,  improve=0.2526931, (0 missing)
      YearRemodAdd < 1983.5  to the left,  improve=0.2157413, (0 missing)
  Surrogate splits:
      YearBuilt    < 1985.5  to the left,   agree=0.826, adj=0.332, (0 split)
      YearRemodAdd < 2002.5  to the left,   agree=0.765, adj=0.097, (0 split)
      GarageArea   < 625.5   to the left,   agree=0.760, adj=0.077, (0 split)
      BsmtFinSF1   < 1333    to the left,   agree=0.743, adj=0.010, (0 split)
      TotRmsAbvGrd splits as  LLLLLLLRLRLL, agree=0.743, adj=0.010, (0 split)

Node number 3: 146 observations,    complexity param=0.05814644
  mean=311708.3, MSE=8.409812e+09
  left son=6 (104 obs) right son=7 (42 obs)
  Primary splits:
      OverallQual  splits as  -------LRR,   improve=0.2856263, (0 missing)
      LotArea      < 12094.5 to the left,   improve=0.2497850, (0 missing)
      TotRmsAbvGrd splits as  --LLLLLLRRR-, improve=0.2481846, (0 missing)
      BsmtFinSF1   < 1224.5  to the left,   improve=0.2341417, (0 missing)
```

```
      GarageArea   < 663     to the left,   improve=0.1742764, (0 missing)
  Surrogate splits:
      BsmtFinSF1   < 1744    to the left,   agree=0.747, adj=0.119, (0 split)
      TotRmsAbvGrd splits as  --LLLLLLLRR-, agree=0.747, adj=0.119, (0 split)
      YearBuilt    < 2007.5  to the left,   agree=0.740, adj=0.095, (0 split)
      LotArea      < 12811.5 to the left,   agree=0.733, adj=0.071, (0 split)
      YearRemodAdd < 2007.5  to the left,   agree=0.733, adj=0.071, (0 split)

Node number 4: 558 observations,    complexity param=0.0291745
  mean=140191.1, MSE=1.416245e+09
  left son=8 (372 obs) right son=9 (186 obs)
  Primary splits:
      FullBath    splits as  LLRR,        improve=0.2226614, (0 missing)
      OverallQual splits as  LLLLLR----,  improve=0.2102913, (0 missing)
      GarageArea  < 387     to the left,  improve=0.1995198, (0 missing)
      Fireplaces  splits as  LRRR,        improve=0.1972087, (0 missing)
      LotArea     < 9100.5  to the left,  improve=0.1645839, (0 missing)
  Surrogate splits:
      TotRmsAbvGrd splits as  LLLLLRRRRRRR, agree=0.781, adj=0.344, (0 split)
      YearBuilt    < 1983.5  to the left,   agree=0.737, adj=0.210, (0 split)
      BedroomAbvGr splits as  LLLLRLRR,     agree=0.729, adj=0.188, (0 split)
      OverallQual  splits as  LLLLLR----,   agree=0.683, adj=0.048, (0 split)
      BsmtFinSF1   < 1106.5  to the left,   agree=0.683, adj=0.048, (0 split)

Node number 5: 196 observations,    complexity param=0.01429023
  mean=209495.3, MSE=2.216673e+09
  left son=10 (174 obs) right son=11 (22 obs)
  Primary splits:
      BsmtFinSF1   < 955.5   to the left,   improve=0.19837900, (0 missing)
      LotArea      < 9701.5  to the left,   improve=0.18976810, (0 missing)
      TotRmsAbvGrd splits as  --LLLLRRRR--, improve=0.18165830, (0 missing)
      GarageArea   < 785     to the left,   improve=0.17263200, (0 missing)
      Fireplaces   splits as  LRRL,         improve=0.08473308, (0 missing)
  Surrogate splits:
      LotArea      < 92955   to the left,  agree=0.898, adj=0.091, (0 split)
      BedroomAbvGr splits as  -RLLLL--,    agree=0.893, adj=0.045, (0 split)

Node number 6: 104 observations,    complexity param=0.01958114
  mean=280562.4, MSE=4.17479e+09
  left son=12 (85 obs) right son=13 (19 obs)
  Primary splits:
      BsmtFinSF1   < 1224.5  to the left,   improve=0.2720096, (0 missing)
      GarageArea   < 536     to the left,   improve=0.2187127, (0 missing)
      LotArea      < 11435.5 to the left,   improve=0.1910548, (0 missing)
      TotRmsAbvGrd splits as  --LLLLLLRRR-, improve=0.1194041, (0 missing)
      BedroomAbvGr splits as  LRLLR---,     improve=0.1144824, (0 missing)
  Surrogate splits:
      LotArea      < 18782.5 to the left,   agree=0.837, adj=0.105, (0 split)
      TotRmsAbvGrd splits as  --LLLLLLLLR-, agree=0.827, adj=0.053, (0 split)
      Fireplaces   splits as  LLLR,         agree=0.827, adj=0.053, (0 split)

Node number 7: 42 observations,    complexity param=0.01801227
  mean=388831.3, MSE=1.05465e+10
  left son=14 (27 obs) right son=15 (15 obs)
```

```
  Primary splits:
      TotRmsAbvGrd splits as  ---LLLLLRRR-, improve=0.2452590, (0 missing)
      GarageArea   < 797     to the left,   improve=0.1844068, (0 missing)
      BsmtFinSF1   < 1277    to the left,   improve=0.1819313, (0 missing)
      LotArea      < 12072   to the left,   improve=0.1793774, (0 missing)
      YearBuilt    < 2007.5  to the left,   improve=0.1734472, (0 missing)
  Surrogate splits:
      BedroomAbvGr splits as  LLLLR---,    agree=0.810, adj=0.467, (0 split)
      Fireplaces   splits as  LLRL,        agree=0.810, adj=0.467, (0 split)
      FullBath     splits as  LLLR,        agree=0.738, adj=0.267, (0 split)
      LotArea      < 18927   to the left,  agree=0.714, adj=0.200, (0 split)
      HalfBath     splits as  LR-,         agree=0.714, adj=0.200, (0 split)

Node number 8: 372 observations,    complexity param=0.0120098
  mean=127634.4, MSE=9.157591e+08
  left son=16 (120 obs) right son=17 (252 obs)
  Primary splits:
      BsmtFinSF1 < 169     to the left,  improve=0.2126306, (0 missing)
      GarageArea < 213     to the left,  improve=0.1896401, (0 missing)
      YearBuilt  < 1952.5  to the left,  improve=0.1737735, (0 missing)
      Fireplaces splits as  LRRR,        improve=0.1733798, (0 missing)
      LotArea    < 9100.5  to the left,  improve=0.1647429, (0 missing)
  Surrogate splits:
      YearBuilt    < 1938.5  to the left,   agree=0.769, adj=0.283, (0 split)
      YearRemodAdd < 1950.5  to the left,   agree=0.742, adj=0.200, (0 split)
      LotArea      < 6411    to the left,   agree=0.702, adj=0.075, (0 split)
      GarageArea   < 230     to the left,   agree=0.691, adj=0.042, (0 split)
      TotRmsAbvGrd splits as  LRRRRRLRL---, agree=0.688, adj=0.033, (0 split)

Node number 9: 186 observations,    complexity param=0.01183168
  mean=165304.6, MSE=1.471188e+09
  left son=18 (64 obs) right son=19 (122 obs)
  Primary splits:
      OverallQual  splits as  --LLLR----,  improve=0.2607831, (0 missing)
      BsmtFinSF1   < 618     to the left,  improve=0.1998511, (0 missing)
      YearRemodAdd < 1980.5  to the left,  improve=0.1856281, (0 missing)
      Fireplaces   splits as  LRR-,        improve=0.1733604, (0 missing)
      LotArea      < 12180   to the left,  improve=0.1715189, (0 missing)
  Surrogate splits:
      YearRemodAdd < 1971.5  to the left,   agree=0.753, adj=0.281, (0 split)
      YearBuilt    < 1971.5  to the left,   agree=0.737, adj=0.234, (0 split)
      TotRmsAbvGrd splits as  --RRRRLRLLLR, agree=0.720, adj=0.188, (0 split)
      GarageArea   < 290     to the left,   agree=0.720, adj=0.188, (0 split)
      BedroomAbvGr splits as  --RRLRLR,     agree=0.704, adj=0.141, (0 split)

Node number 10: 174 observations
  mean=202038.8, MSE=1.600723e+09

Node number 11: 22 observations
  mean=268469.5, MSE=3.17058e+09

Node number 12: 85 observations
  mean=264630.2, MSE=2.666789e+09
```

```
Node number 13: 19 observations
  mean=351838.2, MSE=4.705288e+09

Node number 14: 27 observations
  mean=350923.4, MSE=2.838409e+09

Node number 15: 15 observations
  mean=457065.7, MSE=1.717852e+10

Node number 16: 120 observations
  mean=107412.9, MSE=6.818746e+08

Node number 17: 252 observations
  mean=137263.7, MSE=7.396912e+08

Node number 18: 64 observations
  mean=138261, MSE=1.15381e+09

Node number 19: 122 observations
  mean=179491.4, MSE=1.052756e+09
```

One can notice that the best number of splits to avoid overfitting and get less error is 4.This is because xerror decreases initially and then starts decreasing after 4.So, we have adjusted minsplit value ,s.t, number of splits in our decision tree remains 4.

We observe from the summary that only OverallQual, Garage area, YearBuilt, BsmtfinSF1 are important.Hence, only using these parameters this time to construct the decision tree.

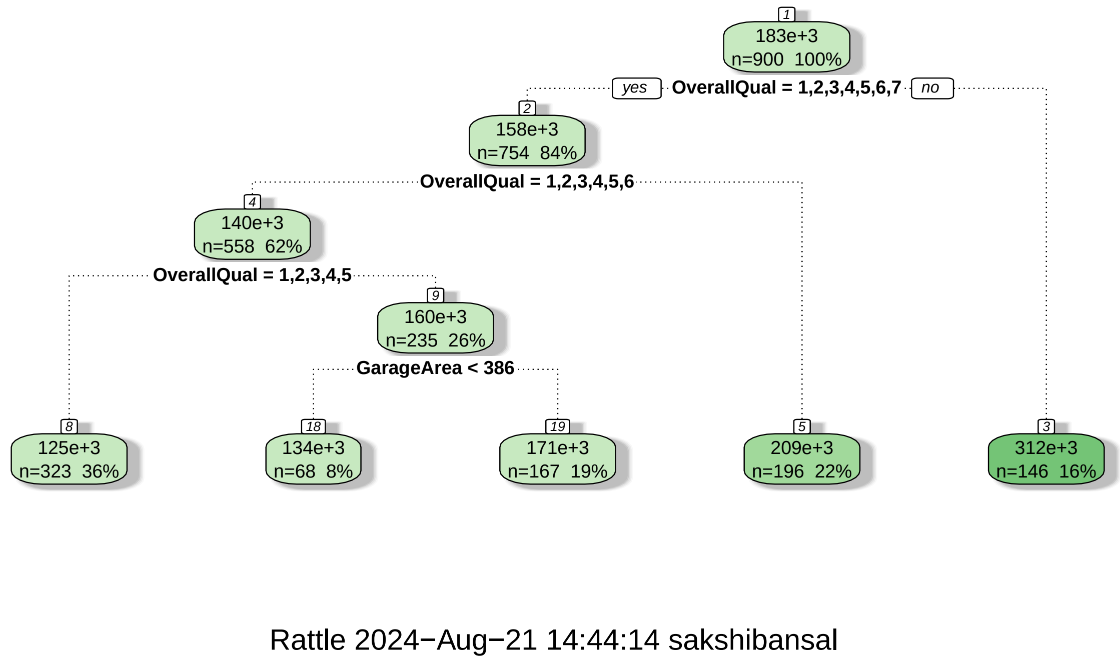
```
d=rpart(SalePrice~OverallQual+GarageArea+YearBuilt+BsmtFinSF1,data = df.houseprice,method = 'anova',con
d$cptable
```

```
          CP nsplit rel error    xerror       xstd
1 0.47785266      0 1.0000000 1.0011350 0.08981504
2 0.11551089      1 0.5221473 0.5251355 0.04671651
3 0.02755368      2 0.4066365 0.4108254 0.04426690
4 0.01122337      3 0.3790828 0.3890547 0.04419121
5 0.01000000      4 0.3678594 0.3834162 0.04417216
```

```
d$variable.importance
```

```
 OverallQual   GarageArea    YearBuilt   BsmtFinSF1
3.744967e+12 1.089056e+12 7.495068e+11 4.019166e+11
```

```
fancyRpartPlot(d)
```



One can notice that the best number of splits to avoid overfitting and get less error is 4.This is because xerror decreases initially and then starts decreasing after 4.So, we have adjusted minsplit value ,s.t, number of splits in our decision tree remains 4.

As now we have only used the significant variables. We can use this model to make predictions in our test dataset.

```
dp=predict(d,df.predict)
head(dp)
```

```
       1        2        3        4        5        6 
125471.0 125471.0 209495.3 209495.3 125471.0 125471.0 
```

Calculating RMSE of predictions:

```
RMSE(dp,df.predict$SalePrice)
```

```
[1] 36441.56
```

```
MAE(dp,df.predict$SalePrice)
```

```
[1] 27999.3
```

Calculating R2 value of the model:

```r
S1 <- sum((dp - df.predict$SalePrice)^2)
S2 <- sum((mean(df.predict$SalePrice) - df.predict$SalePrice)^2)
R <- 1 - (S1 / S2)
round(R,4)
```

```
[1] 0.6414
```

Comparison of linear regression and decision tree:

```r
R_Squared_Value <- c(0.8295, 0.6414)
MAE_Value <- c(18945.79,27999.3)
RMSE_Value <- c(23991.65, 36441.56)

Model <- c("Linear Regression", "Decision Tree")
Model_comparision <- data.frame(Model,R_Squared_Value,RMSE_Value,MAE_Value)
pandoc.table(Model_comparision,style="grid", split.tables = Inf)
```

Model	R_Squared_Value	RMSE_Value	MAE_Value
Linear Regression	0.8295	23992	18946
Decision Tree	0.6414	36442	27999

Interpretation: To determine the most appropriate model for the provided dataset, we assessed two specific models such as linear regression and decision tree. Our evaluation relied on key metrics like R-squared value,RMSE value and MAE value. A preferred model should exhibit a high adjusted R-squared value and a low RMSE/MAE value. Our analysis revealed that the decision tree model had a lower adjusted R-squared value and a higher RMSE/MAE compared to the linear regression model. Consequently, we can conclude that the decision tree model is not suitable for this dataset. Although the linear regression model did not fully satisfy all the assumptions, it demonstrated comparatively better performance than the decision tree model.

Logistic Regression

Classification :

Making a class variable to apply classification on our model. As per the question, Overallqual > 7 is class 1 and rest all is class 0.

```
df.houseprice$OverallQual=as.numeric(df.houseprice$OverallQual)
df.houseprice$OverallQual=as.factor(ifelse(df.houseprice$OverallQual>=7,'1','0'))
df.predict$OverallQual=as.numeric(df.predict$OverallQual)
df.predict$OverallQual=as.factor(ifelse(df.predict$OverallQual>=7,'1','0'))
```

Using logistic regression to predict class:

```
c=glm(OverallQual~.,data = df.houseprice,family = 'binomial')
summary(c)
```

```
Call:
glm(formula = OverallQual ~ ., family = "binomial", data = df.houseprice)

Coefficients: (1 not defined because of singularities)
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -7.202e+01  3.956e+03  -0.018 0.985476
LotArea        -3.673e-05  9.741e-06  -3.771 0.000163 ***
YearBuilt       9.872e-03  6.469e-03   1.526 0.127020
YearRemodAdd    1.747e-02  9.449e-03   1.849 0.064504 .
BsmtFinSF1     -2.146e-03  3.660e-04  -5.862 4.56e-09 ***
FullBath1      -1.694e+00  4.045e+00  -0.419 0.675379
FullBath2      -1.273e+00  4.070e+00  -0.313 0.754464
FullBath3      -9.610e-01  4.250e+00  -0.226 0.821135
HalfBath1      -1.222e-01  2.806e-01  -0.435 0.663267
HalfBath2      -2.249e+00  3.842e+00  -0.585 0.558318
BedroomAbvGr1   1.218e+00  4.019e+00   0.303 0.761929
BedroomAbvGr2   5.267e-01  3.926e+00   0.134 0.893280
BedroomAbvGr3  -2.685e-01  3.951e+00  -0.068 0.945820
BedroomAbvGr4  -6.247e-01  3.983e+00  -0.157 0.875361
BedroomAbvGr5  -1.838e+00  4.281e+00  -0.429 0.667726
BedroomAbvGr6  -1.655e+01  1.589e+03  -0.010 0.991692
BedroomAbvGr8  -6.261e+00  5.595e+03  -0.001 0.999107
TotRmsAbvGrd3  -4.657e+00  4.145e+03  -0.001 0.999104
TotRmsAbvGrd4   9.970e+00  3.956e+03   0.003 0.997989
TotRmsAbvGrd5   1.029e+01  3.956e+03   0.003 0.997926
TotRmsAbvGrd6   1.126e+01  3.956e+03   0.003 0.997729
TotRmsAbvGrd7   1.078e+01  3.956e+03   0.003 0.997827
TotRmsAbvGrd8   1.130e+01  3.956e+03   0.003 0.997720
TotRmsAbvGrd9   1.110e+01  3.956e+03   0.003 0.997761
TotRmsAbvGrd10  1.002e+01  3.956e+03   0.003 0.997979
TotRmsAbvGrd11  1.461e+01  3.956e+03   0.004 0.997054
TotRmsAbvGrd12  1.345e+01  4.117e+03   0.003 0.997393
TotRmsAbvGrd14         NA         NA      NA       NA
Fireplaces1     6.938e-02  2.801e-01   0.248 0.804408
Fireplaces2     5.410e-01  5.064e-01   1.068 0.285354
Fireplaces3     3.287e-01  1.483e+00   0.222 0.824603
GarageArea      1.976e-03  1.061e-03   1.862 0.062544 .
YrSold2007      4.448e-02  3.734e-01   0.119 0.905178
YrSold2008     -1.777e-02  3.741e-01  -0.047 0.962126
YrSold2009     -2.614e-01  3.729e-01  -0.701 0.483298
YrSold2010     -2.418e-01  4.520e-01  -0.535 0.592636
SalePrice       4.576e-05  5.446e-06   8.402  < 2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1195.32  on 899  degrees of freedom
Residual deviance:  455.76  on 864  degrees of freedom
AIC: 527.76

Number of Fisher Scoring iterations: 16
```

Keeping only LotArea, BsmtFinSF1 and SalePrice

```r
c=glm(OverallQual~LotArea+BsmtFinSF1+SalePrice,data = df.houseprice,family = 'binomial')
summary(c)
```

```
Call:
glm(formula = OverallQual ~ LotArea + BsmtFinSF1 + SalePrice,
    family = "binomial", data = df.houseprice)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.359e+00  6.353e-01 -14.733  < 2e-16 ***
LotArea     -4.767e-05  8.954e-06  -5.324 1.01e-07 ***
BsmtFinSF1  -1.879e-03  3.140e-04  -5.983 2.19e-09 ***
SalePrice    5.620e-05  3.894e-06  14.432  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1195.32  on 899  degrees of freedom
Residual deviance:  508.84  on 896  degrees of freedom
AIC: 516.84

Number of Fisher Scoring iterations: 7
```

Now using this model to make predictions on test data

```r
cp=predict(c,df.predict,type = 'response')
cp=ifelse(cp>0.5,1,0)
```

Making confusion matrix of predictions and actual data

```r
C=confusionMatrix(df.predict$OverallQual,as.factor(cp))
C
```

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
```

```
        0 59 21
        1  0 10

              Accuracy : 0.7667
                95% CI : (0.6657, 0.8494)
    No Information Rate : 0.6556
    P-Value [Acc > NIR] : 0.01546

                  Kappa : 0.3844

 Mcnemar's Test P-Value : 1.275e-05

            Sensitivity : 1.0000
            Specificity : 0.3226
         Pos Pred Value : 0.7375
         Neg Pred Value : 1.0000
             Prevalence : 0.6556
         Detection Rate : 0.6556
   Detection Prevalence : 0.8889
      Balanced Accuracy : 0.6613

       'Positive' Class : 0
```

It can be concluded, when logistic regression is applied to the given data to predict categorical variable OverallQual, we get an accuracy of 82.2%, Specifity of 80% and a high precision of 89%. Hence, logistic regression is a robust method to use for prediction of categorical variables in this given dataset.
