Forecasting Future Housing Prices

Mayuraan Cadamban

2023-04-18

Introduction

This project examines data about residential homes in Awes, Iowa, and creates model based off that data to predict the final price of each home using multiple linear regression. Data obtained from https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data.

The data sets have 1460 observations with 79 explanatory variables describing aspects of residential homes in Awes, Iowa.

This report will examine the data using regression analysis, which is a statistical method usd to examine the relationship of two or more variables of interest. The goal is to achieve a regression model with a prediction accuracy above 70%.

```
library(tidyverse)
library(corrplot)
library(ggplot2)
library(lubridate)
library(gridExtra)
library(caTools)
library(GGally)
library(data.table)
library(Matrix)
library(caret)
library(Metrics)
```

Reading the Data

Below, I am reading the data from Kaggle as dataframes into R.

```
train <- read.csv("C:/Users/mc17/Documents/house-prices-advanced-regression-techniques/train.csv")
```

Viewing a Part of the Data

```
head(train,3)
```

```
##   Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 1  1         60       RL          65    8450   Pave  <NA>      Reg         Lvl
```

```
## 2  2         20       RL          80    9600   Pave  <NA>      Reg         Lvl
## 3  3         60       RL          68   11250   Pave  <NA>      IR1         Lvl
##   Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
## 1    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 2    AllPub       FR2       Gtl      Veenker      Feedr       Norm     1Fam
## 3    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
##   HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl
## 1     2Story           7           5      2003         2003     Gable  CompShg
## 2     1Story           6           8      1976         1976     Gable  CompShg
## 3     2Story           7           5      2001         2002     Gable  CompShg
##   Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation
## 1     VinylSd     VinylSd    BrkFace        196        Gd        TA      PConc
## 2     MetalSd     MetalSd       None          0        TA        TA     CBlock
## 3     VinylSd     VinylSd    BrkFace        162        Gd        TA      PConc
##   BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
## 1       Gd       TA           No          GLQ        706          Unf
## 2       Gd       TA           Gd          ALQ        978          Unf
## 3       Gd       TA           Mn          GLQ        486          Unf
##   BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical
## 1          0       150         856    GasA        Ex          Y      SBrkr
## 2          0       284        1262    GasA        Ex          Y      SBrkr
## 3          0       434         920    GasA        Ex          Y      SBrkr
##   X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
## 1       856       854            0      1710            1            0        2
## 2      1262         0            0      1262            0            1        2
## 3       920       866            0      1786            1            0        2
##   HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
## 1        1            3            1          Gd            8        Typ
## 2        0            3            1          TA            6        Typ
## 3        1            3            1          Gd            6        Typ
##   Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
## 1          0        <NA>     Attchd        2003          RFn          2
## 2          1          TA     Attchd        1976          RFn          2
## 3          1          TA     Attchd        2001          RFn          2
##   GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
## 1        548         TA         TA          Y          0          61
## 2        460         TA         TA          Y        298           0
## 3        608         TA         TA          Y          0          42
##   EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature
## 1             0          0           0        0   <NA>  <NA>        <NA>
## 2             0          0           0        0   <NA>  <NA>        <NA>
## 3             0          0           0        0   <NA>  <NA>        <NA>
##   MiscVal MoSold YrSold SaleType SaleCondition SalePrice
## 1       0      2   2008       WD        Normal    208500
## 2       0      5   2007       WD        Normal    181500
## 3       0      9   2008       WD        Normal    223500
```

We can also see the structure and summary statistics of the data, for example the ones for the 'train' data set.

```
str(train)
```

```
## 'data.frame':    1460 obs. of  81 variables:
```

```
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : chr  "RL" "RL" "RL" "RL" ...
##  $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
##  $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
##  $ Street       : chr  "Pave" "Pave" "Pave" "Pave" ...
##  $ Alley        : chr  NA NA NA NA ...
##  $ LotShape     : chr  "Reg" "Reg" "IR1" "IR1" ...
##  $ LandContour  : chr  "Lvl" "Lvl" "Lvl" "Lvl" ...
##  $ Utilities    : chr  "AllPub" "AllPub" "AllPub" "AllPub" ...
##  $ LotConfig    : chr  "Inside" "FR2" "Inside" "Corner" ...
##  $ LandSlope    : chr  "Gtl" "Gtl" "Gtl" "Gtl" ...
##  $ Neighborhood : chr  "CollgCr" "Veenker" "CollgCr" "Crawfor" ...
##  $ Condition1   : chr  "Norm" "Feedr" "Norm" "Norm" ...
##  $ Condition2   : chr  "Norm" "Norm" "Norm" "Norm" ...
##  $ BldgType     : chr  "1Fam" "1Fam" "1Fam" "1Fam" ...
##  $ HouseStyle   : chr  "2Story" "1Story" "2Story" "2Story" ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
##  $ RoofStyle    : chr  "Gable" "Gable" "Gable" "Gable" ...
##  $ RoofMatl     : chr  "CompShg" "CompShg" "CompShg" "CompShg" ...
##  $ Exterior1st  : chr  "VinylSd" "MetalSd" "VinylSd" "Wd Sdng" ...
##  $ Exterior2nd  : chr  "VinylSd" "MetalSd" "VinylSd" "Wd Shng" ...
##  $ MasVnrType   : chr  "BrkFace" "None" "BrkFace" "None" ...
##  $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : chr  "Gd" "TA" "Gd" "TA" ...
##  $ ExterCond    : chr  "TA" "TA" "TA" "TA" ...
##  $ Foundation   : chr  "PConc" "CBlock" "PConc" "BrkTil" ...
##  $ BsmtQual     : chr  "Gd" "Gd" "Gd" "TA" ...
##  $ BsmtCond     : chr  "TA" "TA" "TA" "Gd" ...
##  $ BsmtExposure : chr  "No" "Gd" "Mn" "No" ...
##  $ BsmtFinType1 : chr  "GLQ" "ALQ" "GLQ" "ALQ" ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : chr  "Unf" "Unf" "Unf" "Unf" ...
##  $ BsmtFinSF2   : int  0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF    : int  150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
##  $ Heating      : chr  "GasA" "GasA" "GasA" "GasA" ...
##  $ HeatingQC    : chr  "Ex" "Ex" "Ex" "Gd" ...
##  $ CentralAir   : chr  "Y" "Y" "Y" "Y" ...
##  $ Electrical   : chr  "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...
##  $ X1stFlrSF    : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
##  $ X2ndFlrSF    : int  854 0 866 756 1053 566 0 983 752 0 ...
##  $ LowQualFinSF : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
##  $ BsmtFullBath : int  1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ 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 ...
##  $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual  : chr  "Gd" "TA" "Gd" "Gd" ...
```

```
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : chr  "Typ" "Typ" "Typ" "Typ" ...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : chr  NA "TA" "TA" "Gd" ...
##  $ GarageType   : chr  "Attchd" "Attchd" "Attchd" "Detchd" ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
##  $ GarageFinish : chr  "RFn" "RFn" "RFn" "Unf" ...
##  $ GarageCars   : int  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : chr  "TA" "TA" "TA" "TA" ...
##  $ GarageCond   : chr  "TA" "TA" "TA" "TA" ...
##  $ PavedDrive   : chr  "Y" "Y" "Y" "Y" ...
##  $ WoodDeckSF   : int  0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF  : int  61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch: int  0 0 0 272 0 0 0 228 205 0 ...
##  $ X3SsnPorch   : int  0 0 0 0 0 320 0 0 0 0 ...
##  $ ScreenPorch  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : chr  NA NA NA NA ...
##  $ Fence        : chr  NA NA NA NA ...
##  $ MiscFeature  : chr  NA NA NA NA ...
##  $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
##  $ SaleType     : chr  "WD" "WD" "WD" "WD" ...
##  $ SaleCondition: chr  "Normal" "Normal" "Normal" "Abnorml" ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

```
summary(train)
```

```
##        Id           MSSubClass      MSZoning          LotFrontage
##  Min.   :   1.0   Min.   : 20.0   Length:1460        Min.   : 21.00
##  1st Qu.: 365.8   1st Qu.: 20.0   Class :character   1st Qu.: 59.00
##  Median : 730.5   Median : 50.0   Mode  :character   Median : 69.00
##  Mean   : 730.5   Mean   : 56.9                      Mean   : 70.05
##  3rd Qu.:1095.2   3rd Qu.: 70.0                      3rd Qu.: 80.00
##  Max.   :1460.0   Max.   :190.0                      Max.   :313.00
##                                                      NA's   :259
##     LotArea          Street             Alley             LotShape
##  Min.   :  1300   Length:1460        Length:1460        Length:1460
##  1st Qu.:  7554   Class :character   Class :character   Class :character
##  Median :  9478   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 10517
##  3rd Qu.: 11602
##  Max.   :215245
##
##  LandContour         Utilities          LotConfig          LandSlope
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

```
##  Neighborhood        Condition1         Condition2          BldgType
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   HouseStyle         OverallQual      OverallCond      YearBuilt
##  Length:1460        Min.   : 1.000   Min.   :1.000   Min.   :1872
##  Class :character   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954
##  Mode  :character   Median : 6.000   Median :5.000   Median :1973
##                     Mean   : 6.099   Mean   :5.575   Mean   :1971
##                     3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000
##                     Max.   :10.000   Max.   :9.000   Max.   :2010
##
##   YearRemodAdd   RoofStyle           RoofMatl         Exterior1st
##  Min.   :1950   Length:1460        Length:1460        Length:1460
##  1st Qu.:1967   Class :character   Class :character   Class :character
##  Median :1994   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1985
##  3rd Qu.:2004
##  Max.   :2010
##
##  Exterior2nd         MasVnrType          MasVnrArea      ExterQual
##  Length:1460        Length:1460        Min.   :   0.0   Length:1460
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median :   0.0   Mode  :character
##                                        Mean   : 103.7
##                                        3rd Qu.: 166.0
##                                        Max.   :1600.0
##                                        NA's   :8
##   ExterCond          Foundation          BsmtQual           BsmtCond
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  BsmtExposure       BsmtFinType1         BsmtFinSF1     BsmtFinType2
##  Length:1460        Length:1460        Min.   :   0.0   Length:1460
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median : 383.5   Mode  :character
##                                        Mean   : 443.6
##                                        3rd Qu.: 712.2
##                                        Max.   :5644.0
##
##    BsmtFinSF2        BsmtUnfSF       TotalBsmtSF       Heating
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Length:1460
##  1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8   Class :character
##  Median :   0.00   Median : 477.5   Median : 991.5   Mode  :character
##  Mean   :  46.55   Mean   : 567.2   Mean   :1057.4
##  3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2
```

```
##  Max.   :1474.00   Max.   :2336.0   Max.   :6110.0
##
##   HeatingQC          CentralAir         Electrical          X1stFlrSF
##  Length:1460        Length:1460        Length:1460        Min.   : 334
##  Class :character   Class :character   Class :character   1st Qu.: 882
##  Mode  :character   Mode  :character   Mode  :character   Median :1087
##                                                           Mean   :1163
##                                                           3rd Qu.:1391
##                                                           Max.   :4692
##
##    X2ndFlrSF     LowQualFinSF       GrLivArea     BsmtFullBath
##  Min.   :   0   Min.   :  0.000   Min.   : 334   Min.   :0.0000
##  1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130   1st Qu.:0.0000
##  Median :   0   Median :  0.000   Median :1464   Median :0.0000
##  Mean   : 347   Mean   :  5.845   Mean   :1515   Mean   :0.4253
##  3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777   3rd Qu.:1.0000
##  Max.   :2065   Max.   :572.000   Max.   :5642   Max.   :3.0000
##
##   BsmtHalfBath        FullBath        HalfBath       BedroomAbvGr
##  Min.   :0.00000   Min.   :0.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000
##  Median :0.00000   Median :2.000   Median :0.0000   Median :3.000
##  Mean   :0.05753   Mean   :1.565   Mean   :0.3829   Mean   :2.866
##  3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :2.00000   Max.   :3.000   Max.   :2.0000   Max.   :8.000
##
##   KitchenAbvGr   KitchenQual         TotRmsAbvGrd     Functional
##  Min.   :0.000   Length:1460        Min.   : 2.000   Length:1460
##  1st Qu.:1.000   Class :character   1st Qu.: 5.000   Class :character
##  Median :1.000   Mode  :character   Median : 6.000   Mode  :character
##  Mean   :1.047                      Mean   : 6.518
##  3rd Qu.:1.000                      3rd Qu.: 7.000
##  Max.   :3.000                      Max.   :14.000
##
##    Fireplaces    FireplaceQu         GarageType         GarageYrBlt
##  Min.   :0.000   Length:1460        Length:1460        Min.   :1900
##  1st Qu.:0.000   Class :character   Class :character   1st Qu.:1961
##  Median :1.000   Mode  :character   Mode  :character   Median :1980
##  Mean   :0.613                                         Mean   :1979
##  3rd Qu.:1.000                                         3rd Qu.:2002
##  Max.   :3.000                                         Max.   :2010
##                                                        NA's   :81
##  GarageFinish         GarageCars      GarageArea      GarageQual
##  Length:1460        Min.   :0.000   Min.   :   0.0   Length:1460
##  Class :character   1st Qu.:1.000   1st Qu.: 334.5   Class :character
##  Mode  :character   Median :2.000   Median : 480.0   Mode  :character
##                     Mean   :1.767   Mean   : 473.0
##                     3rd Qu.:2.000   3rd Qu.: 576.0
##                     Max.   :4.000   Max.   :1418.0
##
##   GarageCond         PavedDrive          WoodDeckSF      OpenPorchSF
##  Length:1460        Length:1460        Min.   :  0.00   Min.   :  0.00
##  Class :character   Class :character   1st Qu.:  0.00   1st Qu.:  0.00
##  Mode  :character   Mode  :character   Median :  0.00   Median : 25.00
```

```
##                                     Mean   : 94.24   Mean   : 46.66
##                                     3rd Qu.:168.00   3rd Qu.: 68.00
##                                     Max.   :857.00   Max.   :547.00
##
##  EnclosedPorch      X3SsnPorch      ScreenPorch        PoolArea
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000
##  Median :  0.00   Median :  0.00   Median :  0.00   Median :  0.000
##  Mean   : 21.95   Mean   :  3.41   Mean   : 15.06   Mean   :  2.759
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :552.00   Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##     PoolQC             Fence           MiscFeature          MiscVal
##  Length:1460        Length:1460        Length:1460        Min.   :    0.00
##  Class :character   Class :character   Class :character   1st Qu.:    0.00
##  Mode  :character   Mode  :character   Mode  :character   Median :    0.00
##                                                           Mean   :   43.49
##                                                           3rd Qu.:    0.00
##                                                           Max.   :15500.00
##
##      MoSold           YrSold       SaleType         SaleCondition
##  Min.   : 1.000   Min.   :2006   Length:1460        Length:1460
##  1st Qu.: 5.000   1st Qu.:2007   Class :character   Class :character
##  Median : 6.000   Median :2008   Mode  :character   Mode  :character
##  Mean   : 6.322   Mean   :2008
##  3rd Qu.: 8.000   3rd Qu.:2009
##  Max.   :12.000   Max.   :2010
##
##    SalePrice
##  Min.   : 34900
##  1st Qu.:129975
##  Median :163000
##  Mean   :180921
##  3rd Qu.:214000
##  Max.   :755000
##
```

Now we can check if there are any missing values in the data.

```
NA_values=data.frame(no_of_na_values=colSums(is.na(train))) # checking for null values
head(NA_values,21)
```

```
##              no_of_na_values
## Id                         0
## MSSubClass                 0
## MSZoning                   0
## LotFrontage              259
## LotArea                    0
## Street                     0
## Alley                   1369
## LotShape                   0
## LandContour                0
## Utilities                  0
```

```
## LotConfig                  0
## LandSlope                  0
## Neighborhood               0
## Condition1                 0
## Condition2                 0
## BldgType                   0
## HouseStyle                 0
## OverallQual                0
## OverallCond                0
## YearBuilt                  0
## YearRemodAdd               0
```

As we can see there are some values missing for certain variables as can be seen above, but most variables have 0 missing data points, so there should not be a huge impact in the accuracy of predictions.

Exploratory Data Analysis on Train Data

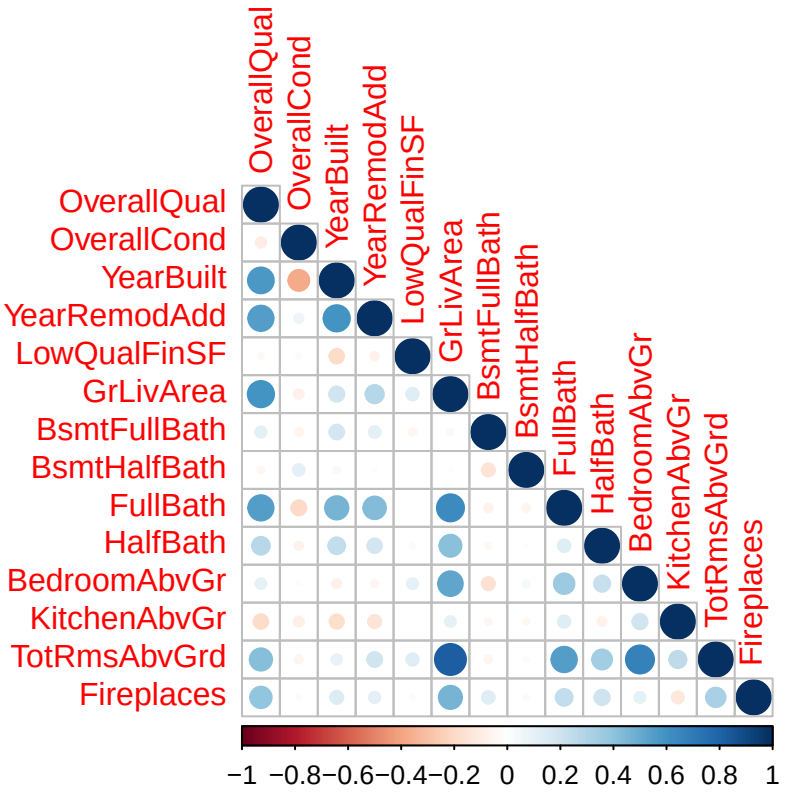
1. Determining Association between Variables

We will create a correlation plot (using the function corrplot) to comprehend the association of the dependent variable (in this case price) with independent variables from the data set.

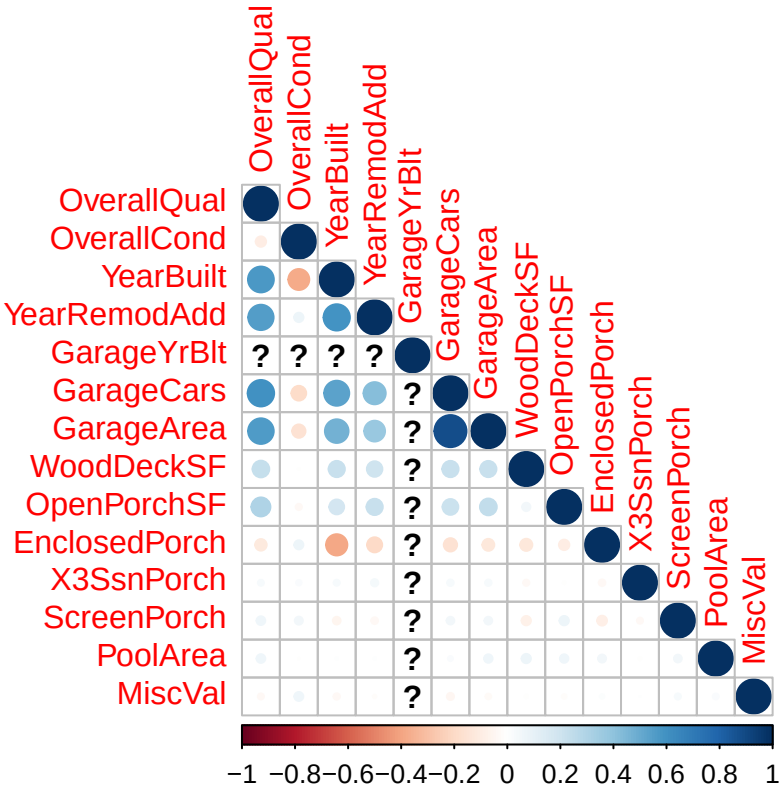
But before doing that we need to drop all the variables that are not numeric so that we can use the variables that can be compared numerically. We are going to split the correlation plots into two plots so we can clearly see the association.

```
train$Street <- NULL # the following variables are not useful in numerical analysis
train$LotShape <- NULL
train$LandContour <- NULL
train$Utilities <- NULL
train$LotConfig <- NULL
train$LandSlope <- NULL
train$Neighborhood <- NULL
train$Condition1 <- NULL
train$Condition2 <- NULL
train$BldgType <- NULL
train$HouseStyle <- NULL
train$RoofStyle <- NULL
train$RoofMatl <- NULL

train$Exterior1st <- NULL
train$Exterior2nd <- NULL
train$MasVnrType <- NULL
train$ExterQual <- NULL
train$ExterCond <- NULL

train$Foundation <- NULL
train$BsmtQual <- NULL
train$BsmtCond <- NULL
train$BsmtExposure <- NULL
train$BsmtFinType1 <- NULL
train$BsmtFinType2 <- NULL
```

```
train$Heating <- NULL
train$HeatingQC <- NULL
train$CentralAir <- NULL
train$Electrical <- NULL
train$KitchenQual <- NULL
train$FireplaceQu <- NULL

train$GarageType <- NULL
train$GarageFinish <- NULL
train$GarageQual <- NULL
train$GarageCond <- NULL
train$PavedDrive <- NULL

train$Functional <- NULL
train$PoolQC <- NULL
train$Fence <- NULL
train$MiscFeature <- NULL
train$SaleType <- NULL
train$SaleCondition <- NULL
train$MSZoning <- NULL
train$Alley <- NULL
correlations <- cor(train[,c(5,6,7,8, 16:25)], use="everything") # first correlation plot
corrplot(correlations, method="circle", type="lower",  sig.level = 0.01, insig = "blank")
```



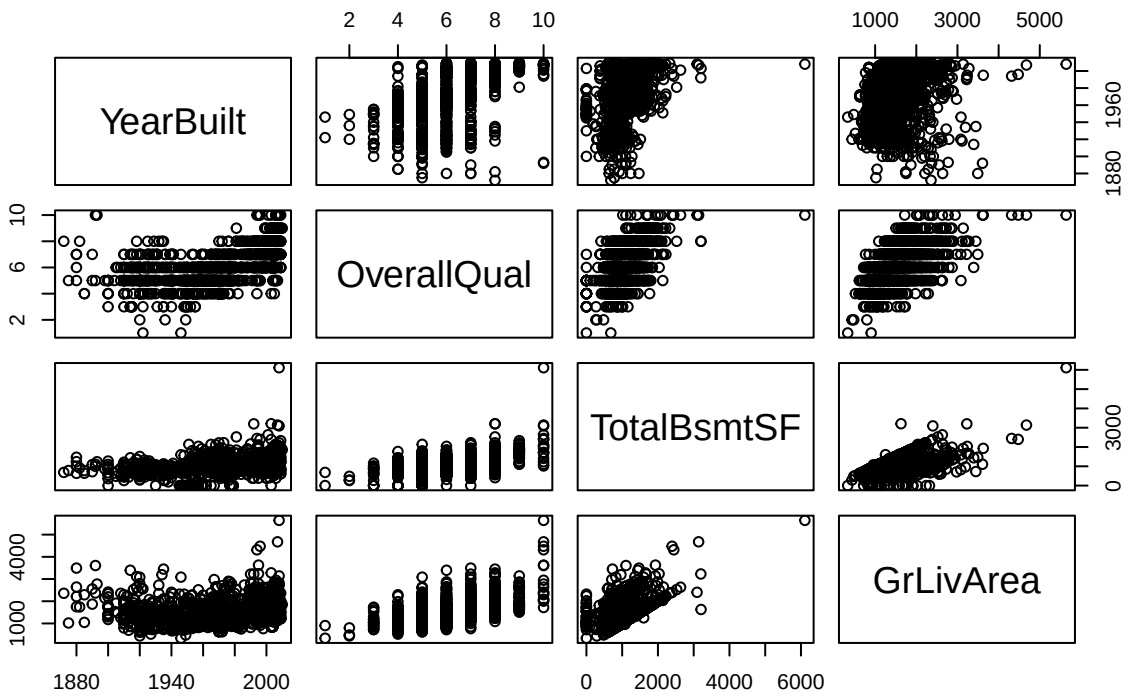
```
correlations <- cor(train[,c(5,6,7,8, 26:35)], use="everything") # second correlation plot
corrplot(correlations, method="circle", type="lower",  sig.level = 0.01, insig = "blank")
```



According to our corrplot, the year a house was built, the amount of garage space and bathrooms was positively correlated to its overall condition, which also contributed to higher sales prices. The rest of the correlations are fairly self - explanatory.

Next we will draw some scatter plots in the form of a matrix to determine the relationship between some of the variables with the strongest correlations. The purpose of putting it in a matrix is so that we can see in a glance how the most important variables are related.

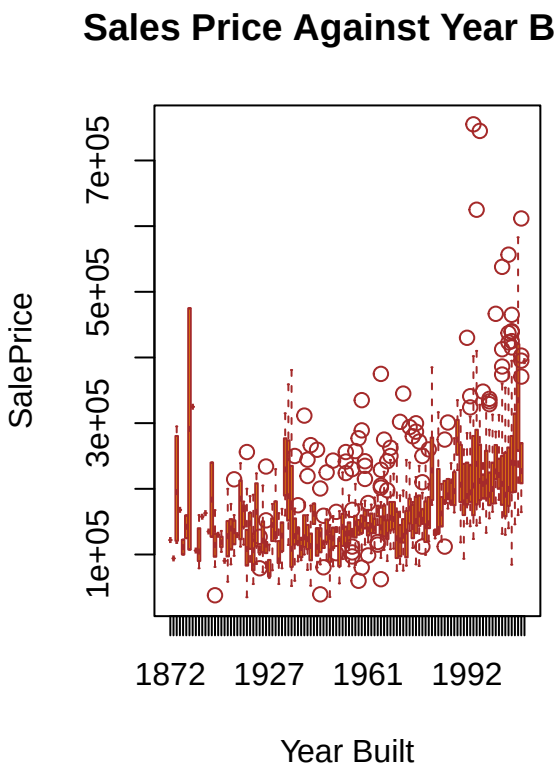
```
pairs(~YearBuilt+OverallQual+TotalBsmtSF+GrLivArea,data=train,
   main="Simple Scatterplot Matrix") # creating a matrix of scatter plots for associated variables
```

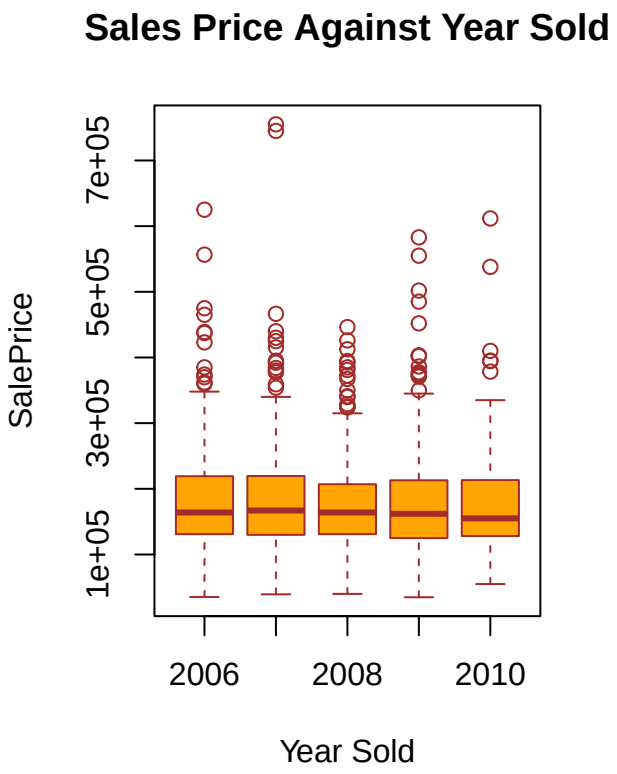


From this we can see that as the years pass by, the total basement square footage has become larger, alongside the size of living areas. It is interesting to see the more square footage is commonly associated with it having better overall quality.

Lets looks at the relation between sales price and the year houses were built/sold.

```
par(mfrow=c(1, 2))
# Box plot of Sales Price Against Year Built
boxplot(SalePrice~YearBuilt,data=train,main="Sales Price Against Year Built", xlab="Year Built",ylab="Sa

# Box plot of Sales Price Against Year Sold
boxplot(SalePrice~YrSold,data=train,main="Sales Price Against Year Sold", xlab="Year Sold",ylab="SalePri
```





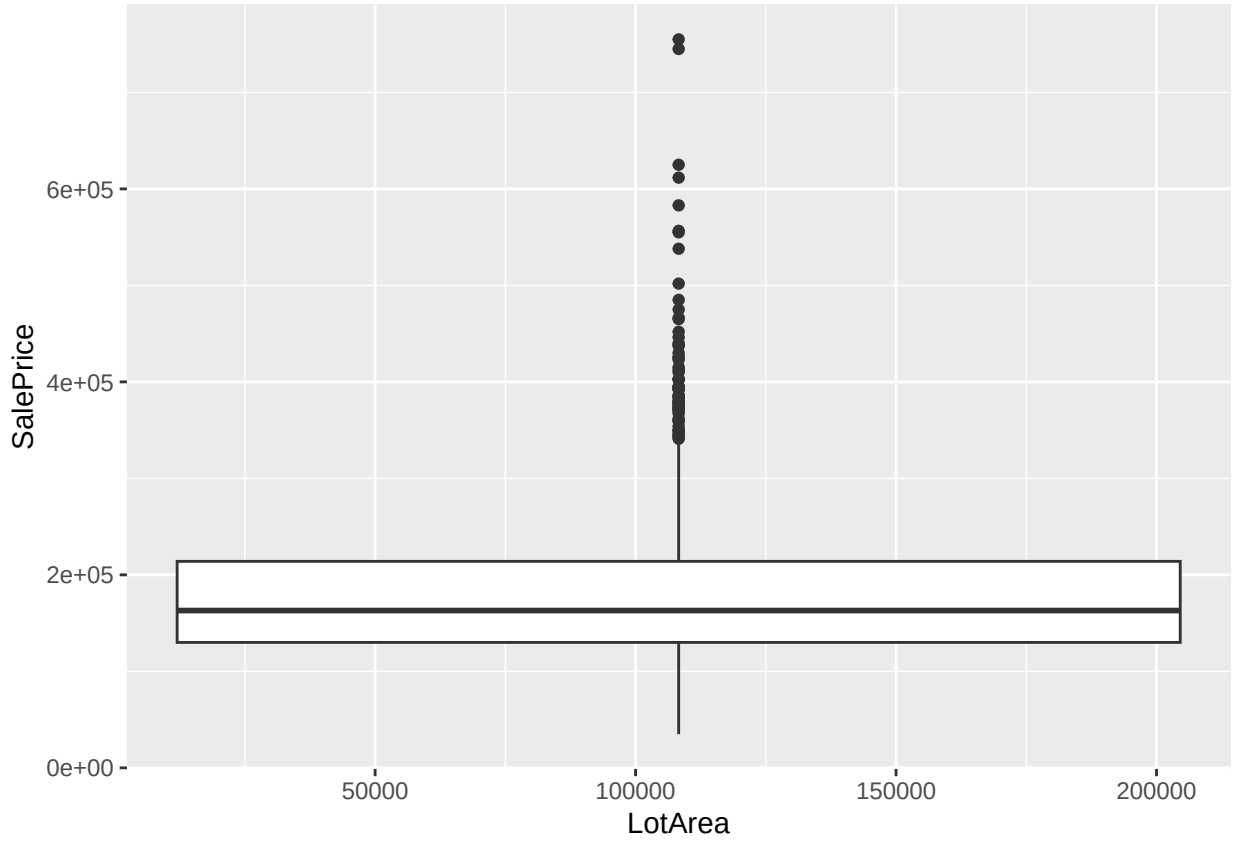
We can see in the box plot "Sales Price Against Year Built" that as time has gone by, the price of new houses built has gone up drastically. Of course this increase is because we are putting sales price against the year built data set, which includes older houses.

In comparison the sales price against years sold shows data only from 2006-2010, so we can see that the sales price has not changed much in these years. But it is interesting to see how different sales prices are now compared to houses built decades ago.

2. Checking for outliers in Dependant variable (Sales Price) using boxplot

Choosing to compare sales price against lot area, as it appears to have a strong correlation.

```
ggplot(data=train)+geom_boxplot(aes(x=LotArea,y=SalePrice))
```



As we can see there is a large number of outliers. We cannot remove these data points as they could be necessary in creating an accurate prediction model.

In order to see how relevant they are, we must compare the fit of a sample linear regression model on the data set with and without outliers.

First we will extract outliers from the data and then obtain the data without the outliers.

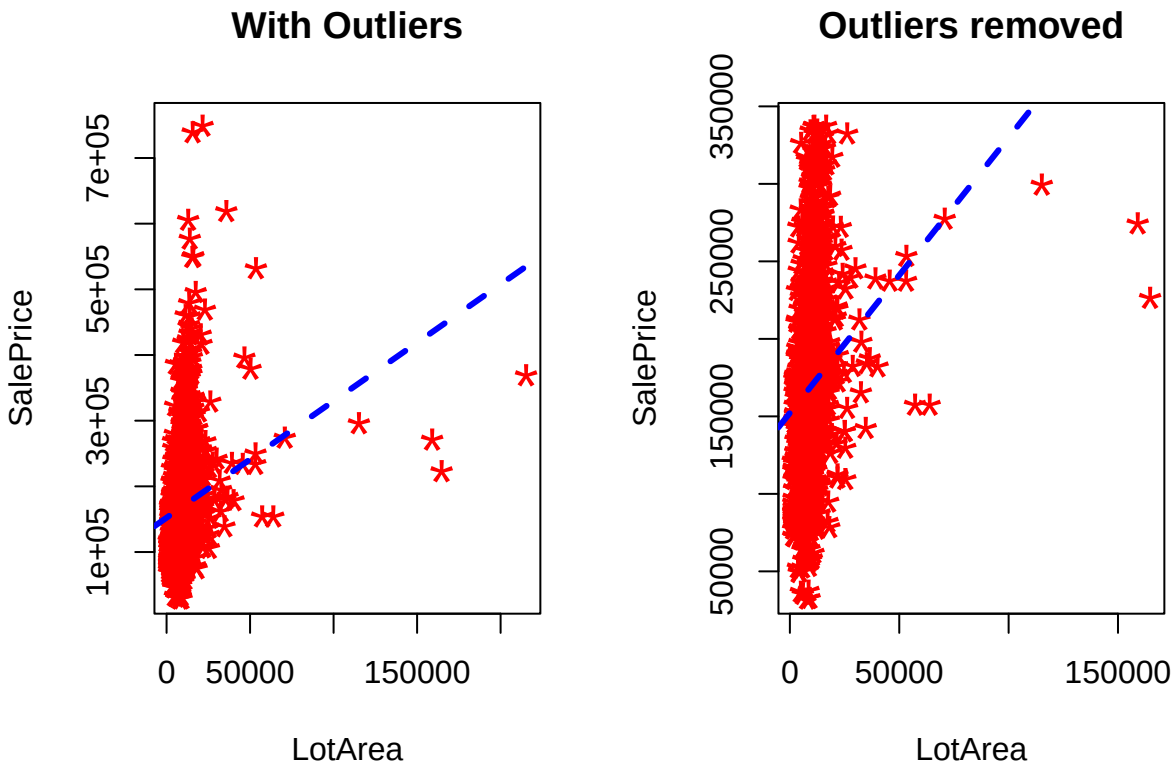
```
outliers=boxplot(train$SalePrice,plot=FALSE)$out # checking for outliers in Train data set for Sale Pri
outliers_data=train[which(train$SalePrice %in% outliers),]
train_data= train[-which(train$SalePrice %in% outliers),]
outliers
```

```
##  [1] 345000 385000 438780 383970 372402 412500 501837 475000 386250 403000
## [11] 415298 360000 375000 342643 354000 377426 437154 394432 426000 555000
## [21] 440000 380000 374000 430000 402861 446261 369900 451950 359100 345000
## [31] 370878 350000 402000 423000 372500 392000 755000 361919 341000 538000
## [41] 395000 485000 582933 385000 350000 611657 395192 348000 556581 424870
## [51] 625000 392500 745000 367294 465000 378500 381000 410000 466500 377500
## [61] 394617
```

We can see there are 61 observations as outliers (which is not that high). Now we can plot the data with and without outliers.

```
par(mfrow=c(1, 2))
# Plot of original data with outliers.
```

```
plot(train$LotArea, train$SalePrice, main="With Outliers", xlab="LotArea", ylab="SalePrice", pch="*", co
abline(lm(SalePrice ~ LotArea, data=train_data), col="blue", lwd=3, lty=2)

# Plot of original data without outliers. We can clearly see a change in slope.
plot(train_data$LotArea, train_data$SalePrice, main="Outliers removed", xlab="LotArea", ylab="SalePrice
abline(lm(SalePrice ~ LotArea, data=train_data), col="blue", lwd=3, lty=2)
```



As we can see above, there is a drastic change in the slope of the best fit line after removing the outliers. There are only 61 outliers, which is quite low looking at the overall data, but those 61 outliers do have a major impact on the model.

Clearly, if we remove the outliers to build our model, our predictions will be exaggerated (high margin of error) for the higher sales price because of the steeper slope.

Now we are ready to build our model.

MODELING

1. Modeling on the entire train data

A linear model using all the variables given in the data set.

```
outcome <- train$SalePrice # first we must partition data to fit model

partition <- createDataPartition(y=outcome,
```

```
                                  p=.5,
                                  list=F)
training <- train[partition,] # partitioning into two sets to create models
testing <- train[-partition,]


lm_model_1 <- lm(SalePrice ~ ., data=training) # generating linear model with all variables
summary(lm_model_1)
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -371446  -19144    -839   16450  289518
##
## Coefficients: (2 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.925e+06  2.495e+06  -0.771 0.440820
## Id             1.695e+00  3.986e+00   0.425 0.670909
## MSSubClass    -2.456e+02  5.237e+01  -4.689 3.50e-06 ***
## LotFrontage   -2.026e+02  8.675e+01  -2.335 0.019902 *
## LotArea        5.553e-01  1.753e-01   3.168 0.001623 **
## OverallQual    1.981e+04  2.187e+03   9.059  < 2e-16 ***
## OverallCond    4.802e+03  2.075e+03   2.314 0.021058 *
## YearBuilt      4.908e+02  1.422e+02   3.451 0.000604 ***
## YearRemodAdd   1.362e+02  1.305e+02   1.044 0.297053
## MasVnrArea     8.415e+00  1.156e+01   0.728 0.467088
## BsmtFinSF1     3.479e+00  8.238e+00   0.422 0.672961
## BsmtFinSF2    -1.032e+01  1.296e+01  -0.797 0.426012
## BsmtUnfSF     -3.502e+00  7.666e+00  -0.457 0.647975
## TotalBsmtSF           NA         NA      NA       NA
## X1stFlrSF      5.701e+01  1.129e+01   5.048 6.16e-07 ***
## X2ndFlrSF      4.516e+01  9.024e+00   5.004 7.65e-07 ***
## LowQualFinSF   1.781e+01  7.780e+01   0.229 0.819052
## GrLivArea             NA         NA      NA       NA
## BsmtFullBath   1.028e+04  4.718e+03   2.178 0.029871 *
## BsmtHalfBath  -4.533e+02  7.077e+03  -0.064 0.948948
## FullBath       4.748e+03  5.351e+03   0.887 0.375291
## HalfBath      -9.691e+02  5.128e+03  -0.189 0.850174
## BedroomAbvGr  -7.726e+03  3.463e+03  -2.231 0.026102 *
## KitchenAbvGr  -2.726e+04  1.035e+04  -2.633 0.008708 **
## TotRmsAbvGrd   6.549e+03  2.254e+03   2.906 0.003818 **
## Fireplaces    -4.788e+02  3.334e+03  -0.144 0.885878
## GarageYrBlt   -2.511e+02  1.443e+02  -1.740 0.082497 .
## GarageCars     2.904e+04  5.240e+03   5.541 4.75e-08 ***
## GarageArea    -3.079e+01  1.781e+01  -1.729 0.084417 .
## WoodDeckSF     3.198e+01  1.454e+01   2.199 0.028288 *
## OpenPorchSF    3.637e+00  2.762e+01   0.132 0.895278
## EnclosedPorch -1.082e+01  3.299e+01  -0.328 0.743136
## X3SsnPorch     1.024e+02  7.013e+01   1.461 0.144745
## ScreenPorch    6.859e+01  2.887e+01   2.376 0.017863 *
```

```
## PoolArea      -1.042e+00  3.781e+01  -0.028 0.978022
## MiscVal       -6.013e+00  1.399e+01  -0.430 0.667563
## MoSold         6.117e+02  6.255e+02   0.978 0.328561
## YrSold         5.511e+02  1.240e+03   0.445 0.656838
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38150 on 526 degrees of freedom
##   (169 observations deleted due to missingness)
## Multiple R-squared:  0.809,  Adjusted R-squared:  0.7963
## F-statistic: 63.64 on 35 and 526 DF,  p-value: < 2.2e-16
```

We can see above that there is an adjusted R-squared value of 0.7567, which is quite high and good for our model, as it indicates approximately 75% of the variation in the outcome is explained using our model.

2. Now we detect the influential points of the data

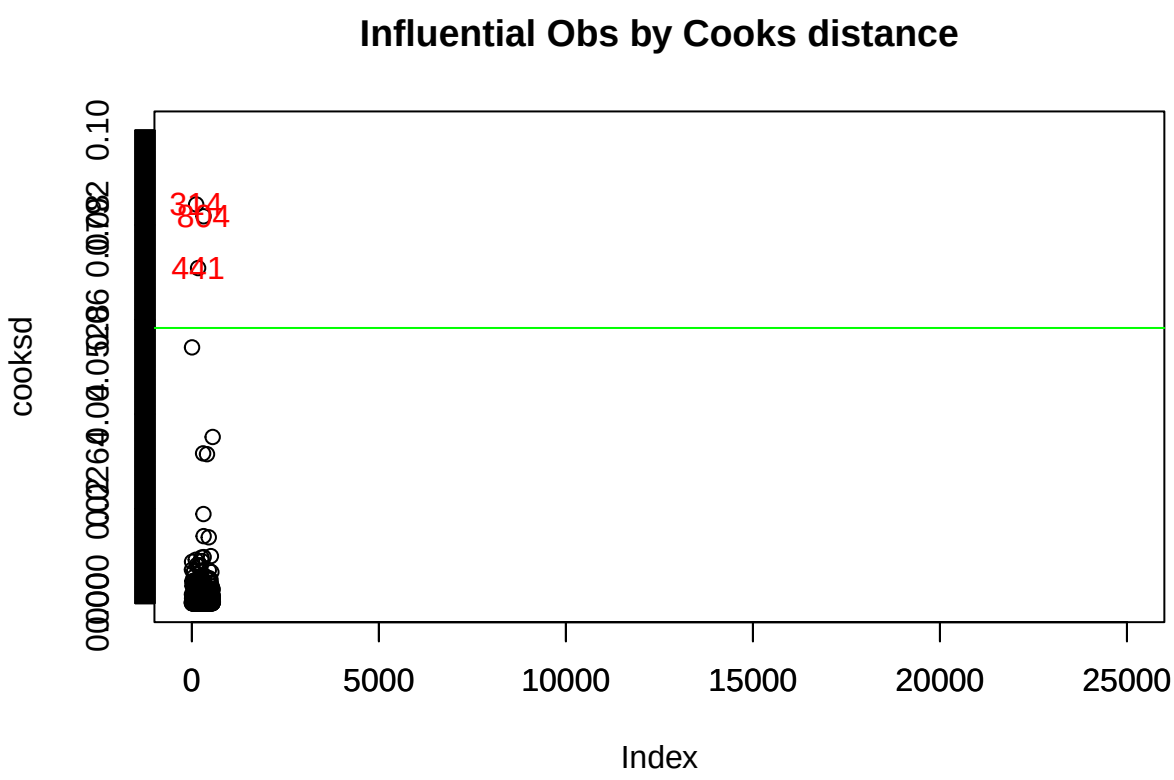
We now must determine the most important observations in our data set. First we determine the cook's distance.

```
cooksd <- cooks.distance(lm_model_1)
mean(cooksd)
```

```
## [1] 0.01455329
```

Now we plot the cook's distance.

```
par(mfrow=c(1, 1))
plot(cooksd, main="Influential Obs by Cooks distance",xlim=c(0,25000),ylim=c(0,0.1))
axis(1, at=seq(0, 25000, 5000))
axis(2, at=seq(0, 0.1, 0.0001))
abline(h = 4*mean(cooksd, na.rm=T), col="green") # line showing where outliers are past relevant data
text(x=1:length(cooksd)+1,y=cooksd,labels=ifelse(cooksd>4*mean(cooksd,na.rm=T),names(cooksd),""), col="
```



Now to find out the influential points in the data.

```
influential <- as.numeric(names(cooksd)[(cooksd > 4*mean(cooksd, na.rm=T))])  # influential row numbers
head(train[influential, ])
```

```
##        Id MSSubClass LotFrontage LotArea OverallQual OverallCond YearBuilt
## 314   314         20         150  215245           7           5      1965
## 441   441         20         105   15431          10           5      2008
## 804   804         60         107   13891           9           5      2008
## 1183 1183         60         160   15623          10           5      1996
## 1299 1299         60         313   63887          10           5      2008
##      YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
## 314          1965          0       1236        820        80        2136
## 441          2008        200       1767        539       788        3094
## 804          2009        424          0          0      1734        1734
## 1183         1996          0       2096          0       300        2396
## 1299         2008        796       5644          0       466        6110
##      X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath
## 314       2036         0            0      2036            2            0
## 441       2402         0            0      2402            1            0
## 804       1734      1088            0      2822            0            0
## 1183      2411      2065            0      4476            1            0
## 1299      4692       950            0      5642            2            0
##      FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces
## 314         2        0            3            1            8          2
## 441         2        0            2            1           10          2
```

```
## 804         3        1            4            1           12          1
## 1183        3        1            4            1           10          2
## 1299        2        1            3            1           12          3
##      GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
## 314         1965          2        513          0           0             0
## 441         2008          3        672          0          72             0
## 804         2009          3       1020         52         170             0
## 1183        1996          3        813        171          78             0
## 1299        2008          2       1418        214         292             0
##      X3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold SalePrice
## 314           0           0        0       0      6   2009    375000
## 441           0         170        0       0      4   2009    555000
## 804           0         192        0       0      1   2009    582933
## 1183          0           0      555       0      7   2007    745000
## 1299          0           0      480       0      1   2008    160000
```

```
influential_data=train[influential, ]
```

Now we take out the influential outliers.

```
influencial_outliers=inner_join(outliers_data,influential_data)
```

```
## Joining with `by = join_by(Id, MSSubClass, LotFrontage, LotArea, OverallQual,
## OverallCond, YearBuilt, YearRemodAdd, MasVnrArea, BsmtFinSF1, BsmtFinSF2,
## BsmtUnfSF, TotalBsmtSF, X1stFlrSF, X2ndFlrSF, LowQualFinSF, GrLivArea,
## BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, BedroomAbvGr, KitchenAbvGr,
## TotRmsAbvGrd, Fireplaces, GarageYrBlt, GarageCars, GarageArea, WoodDeckSF,
## OpenPorchSF, EnclosedPorch, X3SsnPorch, ScreenPorch, PoolArea, MiscVal, MoSold,
## YrSold, SalePrice)`
```

Now we modify the data excluding the outliers and including only the influential outliers.

```
train_data1=rbind(train_data,influencial_outliers)
```

3. Modelling using Train data which includes influential outliers

To create a better model, we will use the modified data which includes influential outliers. We will also try dropping certain variables to see if we can have a better R-squared value.

```
ln_model_2=lm(SalePrice ~ MSSubClass+LotArea+BsmtUnfSF+
                  X1stFlrSF+X2ndFlrSF+GarageCars+
                  WoodDeckSF, data=training ) # select variables and adjusted data
summary(ln_model_2)
```

```
##
## Call:
## lm(formula = SalePrice ~ MSSubClass + LotArea + BsmtUnfSF + X1stFlrSF +
##     X2ndFlrSF + GarageCars + WoodDeckSF, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -422769  -20306   -1561   20615  280324
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.723e+04  6.912e+03  -2.493 0.012876 *
## MSSubClass  -1.557e+02  4.549e+01  -3.423 0.000655 ***
## LotArea      1.661e-01  1.554e-01   1.069 0.285492
## BsmtUnfSF    1.475e+00  4.303e+00   0.343 0.731898
## X1stFlrSF    9.514e+01  5.711e+00  16.659  < 2e-16 ***
## X2ndFlrSF    6.671e+01  4.675e+00  14.270  < 2e-16 ***
## GarageCars   3.606e+04  2.819e+03  12.790  < 2e-16 ***
## WoodDeckSF   7.548e+01  1.451e+01   5.204 2.55e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46880 on 723 degrees of freedom
## Multiple R-squared:  0.6693, Adjusted R-squared:  0.6661
## F-statistic:   209 on 7 and 723 DF,  p-value: < 2.2e-16
```

As we can see in this new model, the adjusted R-squared value of 0.7184 shows that the relationship between the variables shown above to be well interconnected.

As we can see, the R-squared value has not changed too much from the first linear model to the second, showing an equally strong relationship with all the variables. This also means that we did not drop any important variables that would drastically change the results of the model.

Prediction and Accuracy of TRAIN DATA

Now based off our third linear model (ln_model_2) we are ready to predict values and see the accuracy of it. As a reminder we are hoping to achieve an accuracy over 70%.

```
prediction <- predict(ln_model_2, testing, type="response")
model_output <- cbind(testing, prediction)

model_output$log_prediction <- log(model_output$prediction)
model_output$log_SalePrice <- log(model_output$SalePrice)

percentage <- rmse(model_output$log_SalePrice,model_output$log_prediction) # using RMSE to calculate ac
accuracy_test = 1-percentage
accuracy_test
```

```
## [1] 0.7380688
```

We see that the accuracy of the model is approximately 76%.

Thus our model can predict price with an accuracy of approximately 76%.