Final Project Report Appendix

Tam Nguyen

October 28th, 2018

library(caret)  
library(Metrics)

##   
## Attaching package: 'Metrics'

## The following objects are masked from 'package:caret':  
##   
## precision, recall

### Exploring missing value

nrows <- nrow(train) #Numer of rows in train dataset  
missing <- sort(map\_dbl(train, function(x) sum(is.na(x)) / nrows), decreasing = TRUE)  
names\_missing <- names(missing[missing > 0])  
head(missing, 20)

## PoolQC MiscFeature Alley Fence FireplaceQu   
## 0.9965659341 0.9629120879 0.9375000000 0.8076923077 0.4739010989   
## LotFrontage GarageType GarageYrBlt GarageFinish GarageQual   
## 0.1778846154 0.0556318681 0.0556318681 0.0556318681 0.0556318681   
## GarageCond BsmtExposure BsmtFinType2 BsmtQual BsmtCond   
## 0.0556318681 0.0260989011 0.0260989011 0.0254120879 0.0254120879   
## BsmtFinType1 MasVnrType MasVnrArea Electrical Id   
## 0.0254120879 0.0054945055 0.0054945055 0.0006868132 0.0000000000

train$MasVnrType[966] #Checking the NA in MasAreaType

## [1] "BrkFace"

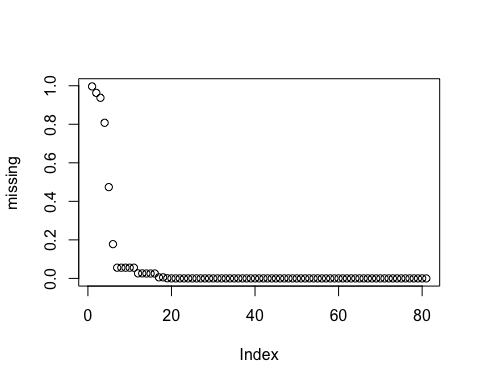
train$MasVnrArea[966] #Checking the coressponding value of MasArea

## [1] 151

summary(missing)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00000 0.00000 0.00000 0.05895 0.00000 0.99657

plot(missing)



* Nineteen (19) of the original 80 variables have some degree of missing values. To begin addressing this problem we work on x\_data for this analysis.

#Make an copy of train dataset to work on 19 variables seprately   
x\_data <- train  
names\_missing\_del <- names(missing[missing > 0.8])  
x\_data <- select(x\_data, one\_of(setdiff(names(x\_data),names\_missing\_del ))) #new training dataset without 4 highest missing predictors

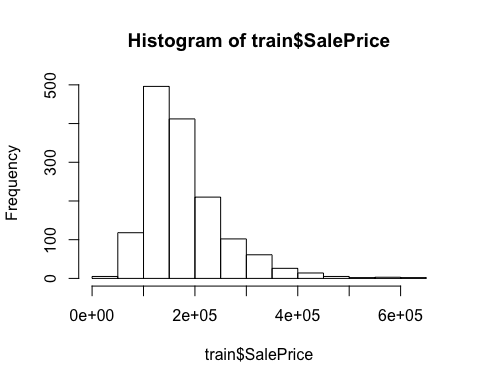
sum(is.na(train)) #how many data is missing

## [1] 6952

sum(is.na(train))/(1460\*80) #missing value percentage

## [1] 0.05952055

hist(train$SalePrice)



### Continuous predictor

For the initial iteration of the problem we first focus on those continuous predictor values. An investigation will be carried out to find good performing models with a focus on identifying (if any) the gap between simple explainable models and the more complex predictive models.

num\_data <- select\_if(x\_data, is.numeric);   
summary(num\_data)

## Id MSSubClass LotFrontage LotArea   
## Min. : 1.0 Min. : 20.00 Min. : 21.00 Min. : 1300   
## 1st Qu.: 364.8 1st Qu.: 20.00 1st Qu.: 59.00 1st Qu.: 7539   
## Median : 730.5 Median : 50.00 Median : 69.00 Median : 9468   
## Mean : 730.0 Mean : 56.89 Mean : 69.69 Mean : 10449   
## 3rd Qu.:1094.2 3rd Qu.: 70.00 3rd Qu.: 80.00 3rd Qu.: 11588   
## Max. :1460.0 Max. :190.00 Max. :313.00 Max. :215245   
## NA's :259   
## OverallQual OverallCond YearBuilt YearRemodAdd   
## Min. : 1.000 Min. :1.000 Min. :1872 Min. :1950   
## 1st Qu.: 5.000 1st Qu.:5.000 1st Qu.:1954 1st Qu.:1967   
## Median : 6.000 Median :5.000 Median :1972 Median :1994   
## Mean : 6.089 Mean :5.576 Mean :1971 Mean :1985   
## 3rd Qu.: 7.000 3rd Qu.:6.000 3rd Qu.:2000 3rd Qu.:2004   
## Max. :10.000 Max. :9.000 Max. :2010 Max. :2010   
##   
## MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF   
## Min. : 0.0 Min. : 0.0 Min. : 0.00 Min. : 0.0   
## 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 0.00 1st Qu.: 222.5   
## Median : 0.0 Median : 381.0 Median : 0.00 Median : 477.5   
## Mean : 102.1 Mean : 437.0 Mean : 46.68 Mean : 567.0   
## 3rd Qu.: 164.2 3rd Qu.: 706.5 3rd Qu.: 0.00 3rd Qu.: 808.0   
## Max. :1600.0 Max. :2188.0 Max. :1474.00 Max. :2336.0   
## NA's :8   
## TotalBsmtSF 1stFlrSF 2ndFlrSF LowQualFinSF   
## Min. : 0.0 Min. : 334 Min. : 0.0 Min. : 0.000   
## 1st Qu.: 795.0 1st Qu.: 882 1st Qu.: 0.0 1st Qu.: 0.000   
## Median : 990.5 Median :1086 Median : 0.0 Median : 0.000   
## Mean :1050.7 Mean :1157 Mean : 343.5 Mean : 5.861   
## 3rd Qu.:1293.8 3rd Qu.:1389 3rd Qu.: 728.0 3rd Qu.: 0.000   
## Max. :3206.0 Max. :3228 Max. :1818.0 Max. :572.000   
##   
## GrLivArea BsmtFullBath BsmtHalfBath FullBath   
## Min. : 334 Min. :0.0000 Min. :0.00000 Min. :0.000   
## 1st Qu.:1128 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:1.000   
## Median :1458 Median :0.0000 Median :0.00000 Median :2.000   
## Mean :1507 Mean :0.4238 Mean :0.05701 Mean :1.562   
## 3rd Qu.:1775 3rd Qu.:1.0000 3rd Qu.:0.00000 3rd Qu.:2.000   
## Max. :3627 Max. :3.0000 Max. :2.00000 Max. :3.000   
##   
## HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd   
## Min. :0.0000 Min. :0.000 Min. :0.000 Min. : 2.000   
## 1st Qu.:0.0000 1st Qu.:2.000 1st Qu.:1.000 1st Qu.: 5.000   
## Median :0.0000 Median :3.000 Median :1.000 Median : 6.000   
## Mean :0.3812 Mean :2.865 Mean :1.047 Mean : 6.506   
## 3rd Qu.:1.0000 3rd Qu.:3.000 3rd Qu.:1.000 3rd Qu.: 7.000   
## Max. :2.0000 Max. :8.000 Max. :3.000 Max. :14.000   
##   
## Fireplaces GarageYrBlt GarageCars GarageArea   
## Min. :0.0000 Min. :1900 Min. :0.000 Min. : 0.0   
## 1st Qu.:0.0000 1st Qu.:1961 1st Qu.:1.000 1st Qu.: 329.5   
## Median :1.0000 Median :1980 Median :2.000 Median : 478.5   
## Mean :0.6092 Mean :1978 Mean :1.764 Mean : 471.6   
## 3rd Qu.:1.0000 3rd Qu.:2002 3rd Qu.:2.000 3rd Qu.: 576.0   
## Max. :3.0000 Max. :2010 Max. :4.000 Max. :1390.0   
## NA's :81   
## WoodDeckSF OpenPorchSF EnclosedPorch 3SsnPorch   
## 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 : 24.00 Median : 0.00 Median : 0.000   
## Mean : 93.83 Mean : 46.22 Mean : 22.01 Mean : 3.419   
## 3rd Qu.:168.00 3rd Qu.: 68.00 3rd Qu.: 0.00 3rd Qu.: 0.000   
## Max. :857.00 Max. :547.00 Max. :552.00 Max. :508.000   
##   
## ScreenPorch PoolArea MiscVal MoSold   
## Min. : 0.0 Min. : 0.000 Min. : 0.00 Min. : 1.000   
## 1st Qu.: 0.0 1st Qu.: 0.000 1st Qu.: 0.00 1st Qu.: 5.000   
## Median : 0.0 Median : 0.000 Median : 0.00 Median : 6.000   
## Mean : 15.1 Mean : 2.056 Mean : 43.61 Mean : 6.326   
## 3rd Qu.: 0.0 3rd Qu.: 0.000 3rd Qu.: 0.00 3rd Qu.: 8.000   
## Max. :480.0 Max. :738.000 Max. :15500.00 Max. :12.000   
##   
## YrSold SalePrice   
## Min. :2006 Min. : 34900   
## 1st Qu.:2007 1st Qu.:129900   
## Median :2008 Median :163000   
## Mean :2008 Mean :180151   
## 3rd Qu.:2009 3rd Qu.:214000   
## Max. :2010 Max. :625000   
##

nrow(num\_data); ncol(num\_data)

## [1] 1456

## [1] 38

### Low variance variables

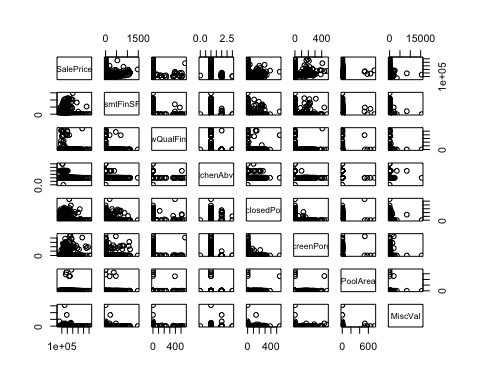
Max Kuhn (2016): Given this, a rule of thumb for detecting near-zero variance predictors is: - The fraction of unique values over the sample size is low (say 10 %). - The ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large (say around 20).

#Function calculating the fraction of unique values over the sample size and the ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value  
condition <- function(x) {  
 checking = list() #emty list  
 tbl = sort(table(x), decreasing = TRUE) #Sorting table decreasing  
 checking[["unique\_to\_samp"]] = length(tbl) / sum(tbl) # Get the variance by divide the length of table to sum of the table  
 checking[["most\_prev\_to\_2nd\_prev"]] = (tbl[[1]] / tbl[[2]]) #get ratio  
 checking  
}  
  
#Function checking if unique\_to\_samp < 0.1 and most\_prev\_to\_2nd\_prev >= 20   
low\_var <- function(x) {  
 low\_var\_vec = vector("character", ncol(x))  
 i = 1  
 for (nme in names(x)) {  
 obs = condition(x[[nme]])  
 #print(obs) #test by printing value  
 if (obs[[1]] <= 0.1 & obs[[2]] >= 20) {   
 low\_var\_vec[i] = nme  
 i = i + 1  
 }  
 }  
 low\_var\_vec[low\_var\_vec != ""]  
}  
  
degen\_vec <- low\_var(num\_data); degen\_vec

## [1] "BsmtFinSF2" "LowQualFinSF" "KitchenAbvGr" "EnclosedPorch"  
## [5] "3SsnPorch" "ScreenPorch" "PoolArea" "MiscVal"

num\_data <- select(num\_data, one\_of(setdiff(names(num\_data), degen\_vec))) #make a new dataset without low variance variables

pairs(~SalePrice + BsmtFinSF2+LowQualFinSF + KitchenAbvGr + EnclosedPorch+ ScreenPorch+PoolArea+MiscVal,data=train)



### Multicolinear

The idea is to first remove the predictors that have the most correlated relationships. - Calculate the correlation matrix of the predictors - Determine the two predictors associated with the largest absolute pairwise correlation (call them predictors A and B). - Determine the average correlation between A and the other variables. Do the same for predictor B. - If A has a larger average correlation, remove it; otherwise, remove predictor B. - Repeat Steps 2–4 until no absolute correlations are above the threshold.

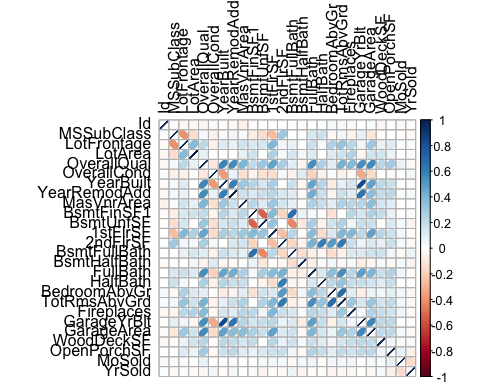
get\_collinear <- function(x) {  
 # Expects data dataframe  
 num\_cols = ncol(x)  
 collinear\_vec = vector("character", num\_cols)   
 index = 1  
   
 for (i in seq(1:num\_cols)) {  
 corMat = cor(x)  
 diag(corMat) = 0 #set diagonal = 0  
 df\_cols = names(x)  
 #Determine the two predictors associated with the largest absolute pairwise correlation (call them predictors A and B).  
 AB = which(corMat == max(abs(corMat), na.rm=TRUE), arr.ind = TRUE)  
 if (corMat[AB][[1]] > 0.75) {  
 names\_AB = rownames(AB)  
   
 if (sum(abs(corMat[names\_AB[1], ]),na.rm=TRUE) > sum(abs(corMat[names\_AB[2], ]),na.rm=TRUE)) {  
   
 collinear\_vec[index] = names\_AB[1]  
 index = index + 1  
 }   
 # if pairwise correlations less than 0.75  
 else {collinear\_vec[index] = names\_AB[2]  
 index = index + 1}  
   
 x = select(x, one\_of(setdiff(df\_cols, collinear\_vec[index - 1])))  
 }  
 else{break}   
 }  
 collinear\_vec[collinear\_vec != ""]  
}  
mul\_col = get\_collinear(num\_data); mul\_col

## [1] "GarageCars" "GrLivArea" "SalePrice" "TotalBsmtSF"

plot(train$GrLivArea, train$SalePrice)



#Correlation matrix for 26 continuous variables  
num\_data <- select(num\_data, one\_of(setdiff(names(num\_data), mul\_col)))  
corrplot(cor(num\_data, use = "pairwise.complete.obs"), method = "ellipse", tl.col = "black", na.label = T)

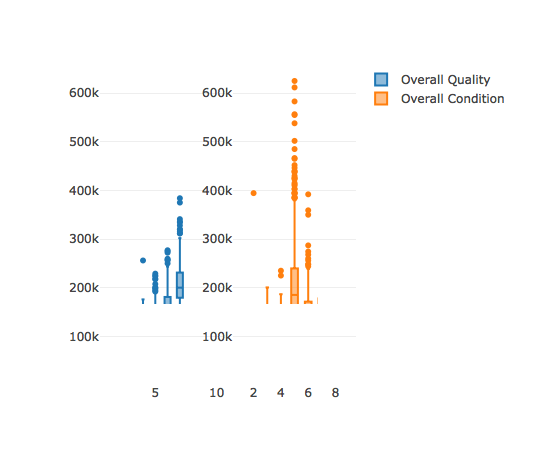


### Decode variables:

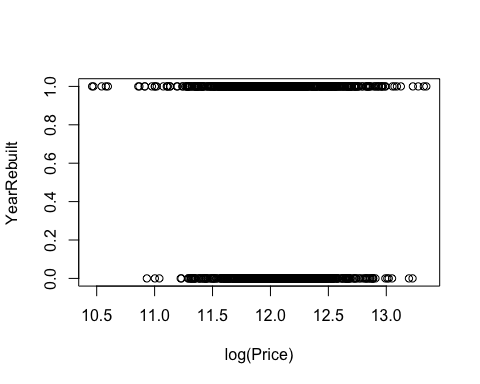
copy\_cont\_var = num\_data #copy the continuous dataset to add more variables  
  
#Checking if having lotshape condition is useable or not  
copy\_cont\_var$LotShape\_new <- ifelse(train$LotShape == 'IR3',0,1)  
  
#Checking if having basement exposure or not  
copy\_cont\_var$BsmtExposure\_new <- ifelse(train$BsmtExposure == 'No',0,1)  
copy\_cont\_var$BsmtExposure\_new[is.na(copy\_cont\_var$BsmtExposure\_new)] = 0 #Change NA value = 0

#Checking if having full bath or halfbath  
copy\_cont\_var$FullBath <- ifelse(train$BsmtFullBath > 0,1,0)  
copy\_cont\_var$HalfBath <- ifelse(train$BsmtHalfBath > 0,1,0)  
  
#Checking if having other Miscellaneous or not  
copy\_cont\_var$MiscFeature\_new = ifelse(train$MiscFeature == 'NA',0,1)  
copy\_cont\_var$MiscFeature\_new[is.na(copy\_cont\_var$MiscFeature\_new)] = 0 #Change NA value = 0  
  
#Checking if having fireplace or not  
copy\_cont\_var$Fireplace = ifelse(train$Fireplaces > 0,1,0)  
  
#Checking if having garage or not  
copy\_cont\_var$GarageYrBlt = ifelse(copy\_cont\_var$GarageYrBlt == 'NA',0,1)  
copy\_cont\_var$GarageYrBlt[is.na(copy\_cont\_var$GarageYrBlt)] = 0 #Change NA value = 0  
  
#Checking if having porch/wood desk ...  
copy\_cont\_var$WoodDeckSF <- as.numeric(copy\_cont\_var$WoodDeckSF)  
copy\_cont\_var$OpenPorchSF <- as.numeric(copy\_cont\_var$OpenPorchSF)  
copy\_cont\_var$Porch = copy\_cont\_var$WoodDeckSF + copy\_cont\_var$OpenPorchSF  
copy\_cont\_var$Porch = ifelse(copy\_cont\_var$Porch > 0, 1, 0) #Change to binary var  
  
#Deleting var  
copy\_cont\_var$WoodDeckSF = NULL  
copy\_cont\_var$OpenPorchSF = NULL  
copy\_cont\_var$Id = NULL  
copy\_cont\_var$LotFrontage = NULL  
copy\_cont\_var$YearBuilt = NULL  
copy\_cont\_var$MoSold = NULL  
copy\_cont\_var$GarageYrBlt = NULL  
copy\_cont\_var$MasVnrArea = NULL  
  
copy\_cont\_var$OverallQual = NULL  
  
#Adding t as new variable for Yearbuild and YearremodAdd  
t = abs(train$YearBuilt - train$YearRemodAdd)  
copy\_cont\_var$YearRebuilt = ifelse(t > 0,1,0)  
copy\_cont\_var$YearRemodAdd = NULL  
  
#Adding Sale Price back to the dataset  
copy\_cont\_var$Price = train$SalePrice

#Relationship between OverallCond vs SalePrice and Overall Quality Condition vs Sale Price  
qual.df <- x\_data[ ,c("OverallQual","OverallCond","SalePrice")]  
pl.q <- plot\_ly(qual.df, y = ~SalePrice, x = ~OverallQual,   
 type = "box", name = "Overall Quality")  
pl.c <- plot\_ly(qual.df, y = ~SalePrice, x = ~OverallCond,   
 type = "box", name = "Overall Condition")  
subplot(pl.q, pl.c)



plot(~ log(Price)+ YearRebuilt, data= copy\_cont\_var)



#Correlation matrix  
cor(copy\_cont\_var[,unlist(lapply(copy\_cont\_var, is.numeric))])

## MSSubClass LotArea OverallCond BsmtFinSF1  
## MSSubClass 1.000000000 -0.142191843 -0.059276572 -0.075268440  
## LotArea -0.142191843 1.000000000 -0.002832285 0.173426158  
## OverallCond -0.059276572 -0.002832285 1.000000000 -0.042542236  
## BsmtFinSF1 -0.075268440 0.173426158 -0.042542236 1.000000000  
## BsmtUnfSF -0.140890171 -0.003774031 -0.137266510 -0.526140244  
## 1stFlrSF -0.265000693 0.267643644 -0.145612855 0.386453075  
## 2ndFlrSF 0.311293638 0.037276582 0.031296654 -0.183357567  
## BsmtFullBath 0.003281653 0.147594611 -0.053106837 0.661932650  
## BsmtHalfBath -0.002508698 0.047390546 0.117206818 0.068868916  
## FullBath -0.009064179 0.094879599 -0.050507518 0.650040531  
## HalfBath -0.009410819 0.051449643 0.124651549 0.070926411  
## BedroomAbvGr -0.023626587 0.118959513 0.013248892 -0.121893063  
## TotRmsAbvGrd 0.040246635 0.173629285 -0.055766348 0.001876651  
## Fireplaces -0.046376588 0.259700916 -0.022277117 0.236218676  
## GarageArea -0.100144776 0.162182789 -0.150679146 0.268650796  
## YrSold -0.021329726 -0.013014088 0.043754812 0.018506484  
## LotShape\_new 0.033524772 -0.227090589 0.056547932 -0.031865941  
## BsmtExposure\_new 0.056941092 0.166805540 -0.048879519 0.291830622  
## MiscFeature\_new -0.041798505 0.111510698 0.074683396 -0.007381151  
## Fireplace -0.034328307 0.178992584 -0.054839297 0.181459637  
## Porch 0.072074637 0.040029685 -0.049925438 0.125974330  
## YearRebuilt -0.058643294 0.005262958 0.308830393 -0.102903548  
## Price -0.088160149 0.269866484 -0.080201802 0.395923108  
## BsmtUnfSF 1stFlrSF 2ndFlrSF BsmtFullBath  
## MSSubClass -0.140890171 -0.265000693 0.311293638 0.003281653  
## LotArea -0.003774031 0.267643644 0.037276582 0.147594611  
## OverallCond -0.137266510 -0.145612855 0.031296654 -0.053106837  
## BsmtFinSF1 -0.526140244 0.386453075 -0.183357567 0.661932650  
## BsmtUnfSF 1.000000000 0.331573791 0.002749242 -0.424026185  
## 1stFlrSF 0.331573791 1.000000000 -0.252296704 0.232826186  
## 2ndFlrSF 0.002749242 -0.252296704 1.000000000 -0.178520522  
## BsmtFullBath -0.424026185 0.232826186 -0.178520522 1.000000000  
## BsmtHalfBath -0.099007488 -0.004382854 -0.032587094 -0.146201453  
## FullBath -0.419570802 0.228352814 -0.175569609 0.977182112  
## HalfBath -0.103517346 -0.005095440 -0.029022283 -0.146709804  
## BedroomAbvGr 0.166583946 0.125474298 0.502450076 -0.152267699  
## TotRmsAbvGrd 0.251935602 0.390639219 0.610793572 -0.063714744  
## Fireplaces 0.051796777 0.396829341 0.182722299 0.130932663  
## GarageArea 0.184562127 0.474245802 0.125023360 0.170653430  
## YrSold -0.040834117 -0.010013637 -0.024874105 0.067665051  
## LotShape\_new 0.009235979 -0.018751917 -0.023791184 -0.003153249  
## BsmtExposure\_new -0.049259417 0.252751320 -0.110433360 0.287506936  
## MiscFeature\_new -0.053132665 -0.048679305 -0.012205931 -0.006206040  
## Fireplace 0.106628919 0.378878579 0.203524492 0.078175318  
## Porch 0.075823949 0.177036500 0.157782114 0.128407160  
## YearRebuilt 0.025787747 -0.020442910 0.102294768 -0.058733294  
## Price 0.220677828 0.625234719 0.297301302 0.235696782  
## BsmtHalfBath FullBath HalfBath BedroomAbvGr  
## MSSubClass -0.002508698 -0.009064179 -0.009410819 -0.023626587  
## LotArea 0.047390546 0.094879599 0.051449643 0.118959513  
## OverallCond 0.117206818 -0.050507518 0.124651549 0.013248892  
## BsmtFinSF1 0.068868916 0.650040531 0.070926411 -0.121893063  
## BsmtUnfSF -0.099007488 -0.419570802 -0.103517346 0.166583946  
## 1stFlrSF -0.004382854 0.228352814 -0.005095440 0.125474298  
## 2ndFlrSF -0.032587094 -0.175569609 -0.029022283 0.502450076  
## BsmtFullBath -0.146201453 0.977182112 -0.146709804 -0.152267699  
## BsmtHalfBath 1.000000000 -0.148245675 0.988073699 0.043330861  
## FullBath -0.148245675 1.000000000 -0.148715136 -0.138015960  
## HalfBath 0.988073699 -0.148715136 1.000000000 0.043942412  
## BedroomAbvGr 0.043330861 -0.138015960 0.043942412 1.000000000  
## TotRmsAbvGrd -0.028715371 -0.063389539 -0.026037151 0.679346237  
## Fireplaces 0.024536785 0.130468632 0.026468455 0.103951004  
## GarageArea -0.028212797 0.184041124 -0.025085131 0.062108286  
## YrSold -0.045302641 0.062771494 -0.041041097 -0.034848689  
## LotShape\_new -0.017951185 0.012725502 -0.019089081 -0.023829596  
## BsmtExposure\_new 0.058827253 0.278445159 0.052258611 -0.104920974  
## MiscFeature\_new 0.029381856 -0.016903413 0.031646877 0.010278126  
## Fireplace 0.030864615 0.083297231 0.032321689 0.103967997  
## Porch 0.022503507 0.130728699 0.026431849 0.022161766  
## YearRebuilt 0.037245185 -0.059955371 0.038345285 0.006575705  
## Price -0.036792474 0.238314540 -0.036647358 0.160541722  
## TotRmsAbvGrd Fireplaces GarageArea YrSold  
## MSSubClass 0.040246635 -0.0463765880 -0.10014478 -0.021329726  
## LotArea 0.173629285 0.2597009161 0.16218279 -0.013014088  
## OverallCond -0.055766348 -0.0222771168 -0.15067915 0.043754812  
## BsmtFinSF1 0.001876651 0.2362186765 0.26865080 0.018506484  
## BsmtUnfSF 0.251935602 0.0517967770 0.18456213 -0.040834117  
## 1stFlrSF 0.390639219 0.3968293413 0.47424580 -0.010013637  
## 2ndFlrSF 0.610793572 0.1827222986 0.12502336 -0.024874105  
## BsmtFullBath -0.063714744 0.1309326632 0.17065343 0.067665051  
## BsmtHalfBath -0.028715371 0.0245367855 -0.02821280 -0.045302641  
## FullBath -0.063389539 0.1304686323 0.18404112 0.062771494  
## HalfBath -0.026037151 0.0264684553 -0.02508513 -0.041041097  
## BedroomAbvGr 0.679346237 0.1039510040 0.06210829 -0.034848689  
## TotRmsAbvGrd 1.000000000 0.3156431699 0.32546680 -0.032189520  
## Fireplaces 0.315643170 1.0000000000 0.25685254 -0.022566883  
## GarageArea 0.325466799 0.2568525448 1.00000000 -0.025870082  
## YrSold -0.032189520 -0.0225668829 -0.02587008 1.000000000  
## LotShape\_new -0.024170881 -0.0481463751 -0.01570848 0.028721352  
## BsmtExposure\_new 0.013371833 0.1443350125 0.23514498 -0.061020307  
## MiscFeature\_new -0.018797916 0.0005849042 -0.03899466 0.057062534  
## Fireplace 0.327188732 0.9032295130 0.30126280 -0.048706535  
## Porch 0.171201959 0.1715273219 0.28074477 -0.001140912  
## YearRebuilt 0.071442881 0.0584579407 -0.12663894 0.021511004  
## Price 0.537461767 0.4667652835 0.63696359 -0.023693833  
## LotShape\_new BsmtExposure\_new MiscFeature\_new Fireplace  
## MSSubClass 0.033524772 0.05694109 -0.0417985050 -0.03432831  
## LotArea -0.227090589 0.16680554 0.1115106984 0.17899258  
## OverallCond 0.056547932 -0.04887952 0.0746833964 -0.05483930  
## BsmtFinSF1 -0.031865941 0.29183062 -0.0073811514 0.18145964  
## BsmtUnfSF 0.009235979 -0.04925942 -0.0531326647 0.10662892  
## 1stFlrSF -0.018751917 0.25275132 -0.0486793047 0.37887858  
## 2ndFlrSF -0.023791184 -0.11043336 -0.0122059310 0.20352449  
## BsmtFullBath -0.003153249 0.28750694 -0.0062060405 0.07817532  
## BsmtHalfBath -0.017951185 0.05882725 0.0293818563 0.03086461  
## FullBath 0.012725502 0.27844516 -0.0169034135 0.08329723  
## HalfBath -0.019089081 0.05225861 0.0316468774 0.03232169  
## BedroomAbvGr -0.023829596 -0.10492097 0.0102781264 0.10396800  
## TotRmsAbvGrd -0.024170881 0.01337183 -0.0187979163 0.32718873  
## Fireplaces -0.048146375 0.14433501 0.0005849042 0.90322951  
## GarageArea -0.015708481 0.23514498 -0.0389946568 0.30126280  
## YrSold 0.028721352 -0.06102031 0.0570625336 -0.04870654  
## LotShape\_new 1.000000000 -0.02115729 0.0154778225 -0.05730133  
## BsmtExposure\_new -0.021157294 1.00000000 -0.0253022724 0.12203426  
## MiscFeature\_new 0.015477823 -0.02530227 1.0000000000 -0.02481544  
## Fireplace -0.057301333 0.12203426 -0.0248154386 1.00000000  
## Porch 0.011135403 0.17517692 -0.0358908424 0.20212124  
## YearRebuilt 0.040174720 -0.09922383 0.0164524932 0.03549190  
## Price -0.043317290 0.31820831 -0.0730224664 0.48054717  
## Porch YearRebuilt Price  
## MSSubClass 0.072074637 -0.058643294 -0.08816015  
## LotArea 0.040029685 0.005262958 0.26986648  
## OverallCond -0.049925438 0.308830393 -0.08020180  
## BsmtFinSF1 0.125974330 -0.102903548 0.39592311  
## BsmtUnfSF 0.075823949 0.025787747 0.22067783  
## 1stFlrSF 0.177036500 -0.020442910 0.62523472  
## 2ndFlrSF 0.157782114 0.102294768 0.29730130  
## BsmtFullBath 0.128407160 -0.058733294 0.23569678  
## BsmtHalfBath 0.022503507 0.037245185 -0.03679247  
## FullBath 0.130728699 -0.059955371 0.23831454  
## HalfBath 0.026431849 0.038345285 -0.03664736  
## BedroomAbvGr 0.022161766 0.006575705 0.16054172  
## TotRmsAbvGrd 0.171201959 0.071442881 0.53746177  
## Fireplaces 0.171527322 0.058457941 0.46676528  
## GarageArea 0.280744765 -0.126638943 0.63696359  
## YrSold -0.001140912 0.021511004 -0.02369383  
## LotShape\_new 0.011135403 0.040174720 -0.04331729  
## BsmtExposure\_new 0.175176923 -0.099223831 0.31820831  
## MiscFeature\_new -0.035890842 0.016452493 -0.07302247  
## Fireplace 0.202121243 0.035491902 0.48054717  
## Porch 1.000000000 -0.062525470 0.34801088  
## YearRebuilt -0.062525470 1.000000000 -0.02356157  
## Price 0.348010882 -0.023561568 1.00000000

#Continuing deleting variable by intuition  
copy\_2 = copy\_cont\_var  
copy\_2$Porch = NULL  
copy\_2$MiscFeature\_new = NULL  
copy\_2$GrLivArea = NULL  
copy\_2$YrSold = NULL  
copy\_2$GarageYrBlt = NULL  
copy\_2$BsmtHalfBath = NULL  
copy\_2$HalfBath = NULL  
copy\_2$Fireplaces = NULL  
copy\_2$BsmtFullBath = NULL  
copy\_2$TotRmsAbvGrd = NULL  
copy\_2$YearRemod = NULL  
copy\_2$BedroomAbvGr = NULL   
copy\_2$LotShape\_new = NULL  
copy\_2$MasVnrArea = NULL  
  
copy\_2$MSSubClass = as.factor(copy\_2$MSSubClass) #Add MSSubClass as an indicator to the dataset  
  
#copy\_2$MasVnrArea = as.factor(ifelse(train$MasVnrArea >0,1,0))  
#copy\_2$newTotalArea = copy\_cont\_var$GrLivArea + train$TotalBsmtSF

### Building Model

#Base line model with all variables (housing assesment and location)  
baseline = lm(copy\_2$Price~. +train$Neighborhood, data = copy\_2);  
summary(baseline); anova(baseline)

##   
## Call:  
## lm(formula = copy\_2$Price ~ . + train$Neighborhood, data = copy\_2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -135449 -14924 407 13320 197956   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.556e+02 1.059e+04 -0.043 0.965707   
## MSSubClass30 -1.182e+04 4.803e+03 -2.462 0.013954 \*   
## MSSubClass40 -2.675e+03 1.489e+04 -0.180 0.857481   
## MSSubClass45 -9.127e+02 8.986e+03 -0.102 0.919107   
## MSSubClass50 -1.615e+04 4.295e+03 -3.760 0.000177 \*\*\*  
## MSSubClass60 5.271e+03 4.794e+03 1.099 0.271775   
## MSSubClass70 -1.920e+04 6.051e+03 -3.173 0.001540 \*\*   
## MSSubClass75 -1.104e+04 9.063e+03 -1.218 0.223595   
## MSSubClass80 -4.918e+03 4.368e+03 -1.126 0.260370   
## MSSubClass85 -7.084e+03 6.966e+03 -1.017 0.309407   
## MSSubClass90 -2.974e+04 4.652e+03 -6.392 2.22e-10 \*\*\*  
## MSSubClass120 -1.634e+04 4.169e+03 -3.919 9.33e-05 \*\*\*  
## MSSubClass160 -2.906e+04 6.524e+03 -4.454 9.10e-06 \*\*\*  
## MSSubClass180 -4.201e+03 1.170e+04 -0.359 0.719518   
## MSSubClass190 -3.915e+04 6.498e+03 -6.026 2.15e-09 \*\*\*  
## LotArea 4.372e-01 9.045e-02 4.834 1.49e-06 \*\*\*  
## OverallCond 7.161e+03 8.000e+02 8.950 < 2e-16 \*\*\*  
## BsmtFinSF1 4.149e+01 3.420e+00 12.132 < 2e-16 \*\*\*  
## BsmtUnfSF 2.315e+01 3.135e+00 7.385 2.61e-13 \*\*\*  
## `1stFlrSF` 7.518e+01 3.935e+00 19.106 < 2e-16 \*\*\*  
## `2ndFlrSF` 6.562e+01 4.349e+00 15.088 < 2e-16 \*\*\*  
## FullBath 4.717e+03 2.155e+03 2.189 0.028751 \*   
## GarageArea 3.584e+01 4.973e+00 7.206 9.34e-13 \*\*\*  
## BsmtExposure\_new 1.247e+04 1.975e+03 6.316 3.60e-10 \*\*\*  
## Fireplace 5.145e+03 1.967e+03 2.616 0.008987 \*\*   
## YearRebuilt 1.982e+03 1.901e+03 1.043 0.297282   
## train$NeighborhoodBlueste -9.073e+03 2.279e+04 -0.398 0.690665   
## train$NeighborhoodBrDale -1.564e+04 1.199e+04 -1.304 0.192368   
## train$NeighborhoodBrkSide -3.369e+04 9.572e+03 -3.520 0.000445 \*\*\*  
## train$NeighborhoodClearCr -3.510e+04 1.002e+04 -3.502 0.000476 \*\*\*  
## train$NeighborhoodCollgCr -1.829e+04 8.422e+03 -2.172 0.030057 \*   
## train$NeighborhoodCrawfor -1.343e+04 9.423e+03 -1.425 0.154316   
## train$NeighborhoodEdwards -4.121e+04 8.918e+03 -4.621 4.18e-06 \*\*\*  
## train$NeighborhoodGilbert -1.786e+04 8.888e+03 -2.010 0.044663 \*   
## train$NeighborhoodIDOTRR -4.550e+04 9.986e+03 -4.556 5.66e-06 \*\*\*  
## train$NeighborhoodMeadowV -4.070e+04 1.199e+04 -3.394 0.000708 \*\*\*  
## train$NeighborhoodMitchel -3.925e+04 9.205e+03 -4.264 2.14e-05 \*\*\*  
## train$NeighborhoodNAmes -4.633e+04 8.541e+03 -5.424 6.86e-08 \*\*\*  
## train$NeighborhoodNoRidge 8.894e+03 9.556e+03 0.931 0.352205   
## train$NeighborhoodNPkVill -1.444e+04 1.263e+04 -1.144 0.252949   
## train$NeighborhoodNridgHt 4.720e+04 8.406e+03 5.615 2.36e-08 \*\*\*  
## train$NeighborhoodNWAmes -4.672e+04 8.906e+03 -5.246 1.79e-07 \*\*\*  
## train$NeighborhoodOldTown -4.884e+04 9.130e+03 -5.349 1.03e-07 \*\*\*  
## train$NeighborhoodSawyer -4.674e+04 9.040e+03 -5.170 2.68e-07 \*\*\*  
## train$NeighborhoodSawyerW -2.612e+04 8.936e+03 -2.923 0.003525 \*\*   
## train$NeighborhoodSomerst 1.455e+04 8.642e+03 1.684 0.092427 .   
## train$NeighborhoodStoneBr 5.223e+04 9.549e+03 5.470 5.32e-08 \*\*\*  
## train$NeighborhoodSWISU -4.556e+04 1.059e+04 -4.302 1.81e-05 \*\*\*  
## train$NeighborhoodTimber -1.433e+04 9.496e+03 -1.510 0.131391   
## train$NeighborhoodVeenker 8.784e+02 1.182e+04 0.074 0.940791   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 29190 on 1406 degrees of freedom  
## Multiple R-squared: 0.86, Adjusted R-squared: 0.8551   
## F-statistic: 176.2 on 49 and 1406 DF, p-value: < 2.2e-16

## Analysis of Variance Table  
##   
## Response: copy\_2$Price  
## Df Sum Sq Mean Sq F value Pr(>F)   
## MSSubClass 14 2.1514e+12 1.5367e+11 180.3001 < 2.2e-16 \*\*\*  
## LotArea 1 4.4644e+11 4.4644e+11 523.7885 < 2.2e-16 \*\*\*  
## OverallCond 1 1.6247e+08 1.6247e+08 0.1906 0.6624699   
## BsmtFinSF1 1 8.3256e+11 8.3256e+11 976.8105 < 2.2e-16 \*\*\*  
## BsmtUnfSF 1 1.9343e+12 1.9343e+12 2269.4616 < 2.2e-16 \*\*\*  
## `1stFlrSF` 1 8.2572e+11 8.2572e+11 968.7903 < 2.2e-16 \*\*\*  
## `2ndFlrSF` 1 2.8878e+11 2.8878e+11 338.8175 < 2.2e-16 \*\*\*  
## FullBath 1 1.6062e+10 1.6062e+10 18.8447 1.52e-05 \*\*\*  
## GarageArea 1 1.9229e+11 1.9229e+11 225.6038 < 2.2e-16 \*\*\*  
## BsmtExposure\_new 1 8.9942e+10 8.9942e+10 105.5256 < 2.2e-16 \*\*\*  
## Fireplace 1 1.1864e+10 1.1864e+10 13.9196 0.0001984 \*\*\*  
## YearRebuilt 1 8.3612e+09 8.3612e+09 9.8098 0.0017716 \*\*   
## train$Neighborhood 24 5.6254e+11 2.3439e+10 27.5005 < 2.2e-16 \*\*\*  
## Residuals 1406 1.1984e+12 8.5232e+08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

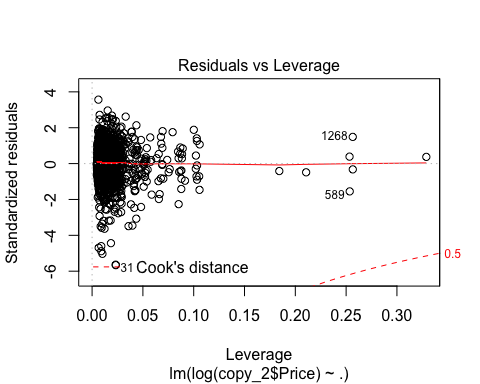
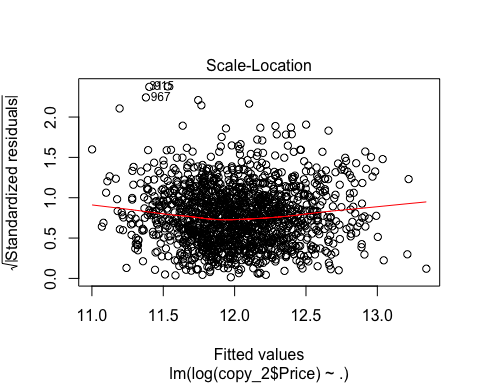
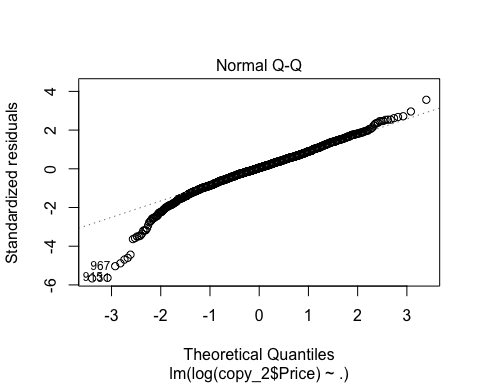
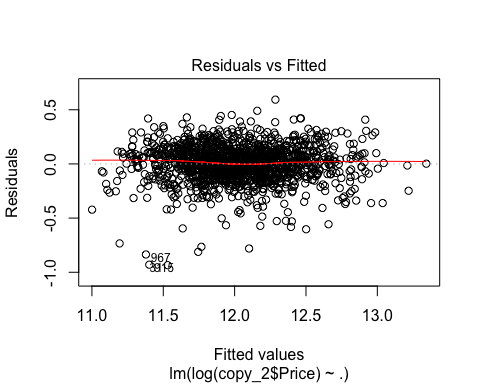
* With all variables related to housing assesment and housing location, the baseline model was created with an ajusted R-squared is 85.51%

## Model with housing assesment only

# Model with Year Built

#Adding t as new variable for Yearbuild and YearremodAdd  
t = abs(train$YearBuilt - train$YearRemodAdd)  
copy\_2$YearRebuilt = ifelse(t>0,1,0)  
model1.1 = lm(log(copy\_2$Price) ~., data = copy\_2); summary(model1.1); plot(model1.1); anova(model1.1)

##   
## Call:  
## lm(formula = log(copy\_2$Price) ~ ., data = copy\_2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.93439 -0.08879 0.00827 0.10031 0.59340   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.073e+01 3.393e-02 316.212 < 2e-16 \*\*\*  
## MSSubClass30 -2.248e-01 2.361e-02 -9.519 < 2e-16 \*\*\*  
## MSSubClass40 -8.642e-02 8.427e-02 -1.025 0.305309   
## MSSubClass45 -8.680e-02 4.966e-02 -1.748 0.080683 .   
## MSSubClass50 -1.212e-01 2.197e-02 -5.517 4.10e-08 \*\*\*  
## MSSubClass60 1.106e-01 2.650e-02 4.174 3.18e-05 \*\*\*  
## MSSubClass70 -9.394e-02 3.141e-02 -2.991 0.002828 \*\*   
## MSSubClass75 -1.513e-01 4.901e-02 -3.087 0.002062 \*\*   
## MSSubClass80 -1.022e-02 2.467e-02 -0.414 0.678720   
## MSSubClass85 -4.217e-02 3.948e-02 -1.068 0.285634   
## MSSubClass90 -1.800e-01 2.605e-02 -6.908 7.37e-12 \*\*\*  
## MSSubClass120 1.174e-01 2.003e-02 5.863 5.65e-09 \*\*\*  
## MSSubClass160 1.526e-02 2.956e-02 0.516 0.605688   
## MSSubClass180 -1.327e-01 5.539e-02 -2.396 0.016701 \*   
## MSSubClass190 -2.498e-01 3.468e-02 -7.202 9.59e-13 \*\*\*  
## LotArea 1.726e-06 4.832e-07 3.572 0.000366 \*\*\*  
## OverallCond 4.871e-02 4.367e-03 11.153 < 2e-16 \*\*\*  
## BsmtFinSF1 2.532e-04 1.905e-05 13.289 < 2e-16 \*\*\*  
## BsmtUnfSF 2.232e-04 1.701e-05 13.119 < 2e-16 \*\*\*  
## `1stFlrSF` 3.615e-04 2.172e-05 16.648 < 2e-16 \*\*\*  
## `2ndFlrSF` 3.276e-04 2.420e-05 13.536 < 2e-16 \*\*\*  
## FullBath 4.611e-02 1.216e-02 3.791 0.000156 \*\*\*  
## GarageArea 3.655e-04 2.660e-05 13.740 < 2e-16 \*\*\*  
## BsmtExposure\_new 7.365e-02 1.074e-02 6.859 1.03e-11 \*\*\*  
## Fireplace 6.408e-02 1.066e-02 6.010 2.36e-09 \*\*\*  
## YearRebuilt -4.531e-03 1.064e-02 -0.426 0.670276   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1669 on 1430 degrees of freedom  
## Multiple R-squared: 0.8254, Adjusted R-squared: 0.8224   
## F-statistic: 270.4 on 25 and 1430 DF, p-value: < 2.2e-16



## Analysis of Variance Table  
##   
## Response: log(copy\_2$Price)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## MSSubClass 14 75.318 5.380 193.0464 < 2.2e-16 \*\*\*  
## LotArea 1 10.283 10.283 368.9758 < 2.2e-16 \*\*\*  
## OverallCond 1 0.586 0.586 21.0291 4.919e-06 \*\*\*  
## BsmtFinSF1 1 17.891 17.891 641.9693 < 2.2e-16 \*\*\*  
## BsmtUnfSF 1 49.561 49.561 1778.3891 < 2.2e-16 \*\*\*  
## `1stFlrSF` 1 19.636 19.636 704.6115 < 2.2e-16 \*\*\*  
## `2ndFlrSF` 1 6.249 6.249 224.2394 < 2.2e-16 \*\*\*  
## FullBath 1 0.726 0.726 26.0395 3.794e-07 \*\*\*  
## GarageArea 1 5.844 5.844 209.7041 < 2.2e-16 \*\*\*  
## BsmtExposure\_new 1 1.306 1.306 46.8730 1.121e-11 \*\*\*  
## Fireplace 1 1.002 1.002 35.9623 2.544e-09 \*\*\*  
## YearRebuilt 1 0.005 0.005 0.1814 0.6703   
## Residuals 1430 39.852 0.028   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* Adjusted R-squared: 0.8224
* MSE: 0.028
* Year rebuilt is not significant

# Recode basement area var

newbsmt = train$TotalBsmtSF - train$BsmtUnfSF  
model.nbsmt = lm(Price ~ newbsmt+ LotArea+ `1stFlrSF`+ `2ndFlrSF`+ FullBath+GarageArea+BsmtExposure\_new+Fireplace +OverallCond+MSSubClass + YearRebuilt,data = copy\_2 ); summary(model.nbsmt)

##   
## Call:  
## lm(formula = Price ~ newbsmt + LotArea + `1stFlrSF` + `2ndFlrSF` +   
## FullBath + GarageArea + BsmtExposure\_new + Fireplace + OverallCond +   
## MSSubClass + YearRebuilt, data = copy\_2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -157462 -18372 -664 17983 245701   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.787e+04 7.513e+03 -6.372 2.51e-10 \*\*\*  
## newbsmt 1.678e+01 3.237e+00 5.185 2.47e-07 \*\*\*  
## LotArea 2.790e-01 1.077e-01 2.591 0.009656 \*\*   
## `1stFlrSF` 1.083e+02 4.059e+00 26.672 < 2e-16 \*\*\*  
## `2ndFlrSF` 7.149e+01 5.384e+00 13.279 < 2e-16 \*\*\*  
## FullBath 4.176e+03 2.772e+03 1.507 0.132098   
## GarageArea 7.630e+01 5.856e+00 13.029 < 2e-16 \*\*\*  
## BsmtExposure\_new 2.159e+04 2.386e+03 9.051 < 2e-16 \*\*\*  
## Fireplace 5.984e+03 2.370e+03 2.525 0.011668 \*   
## OverallCond 3.729e+03 9.687e+02 3.849 0.000124 \*\*\*  
## MSSubClass30 -1.182e+04 5.259e+03 -2.248 0.024751 \*   
## MSSubClass40 -5.157e+03 1.874e+04 -0.275 0.783198   
## MSSubClass45 6.159e+03 1.105e+04 0.558 0.577176   
## MSSubClass50 -2.149e+04 4.890e+03 -4.395 1.19e-05 \*\*\*  
## MSSubClass60 1.518e+04 5.894e+03 2.576 0.010107 \*   
## MSSubClass70 -2.064e+04 6.995e+03 -2.951 0.003218 \*\*   
## MSSubClass75 -2.968e+04 1.091e+04 -2.721 0.006580 \*\*   
## MSSubClass80 -2.157e+04 5.418e+03 -3.980 7.23e-05 \*\*\*  
## MSSubClass85 -1.838e+04 8.770e+03 -2.096 0.036268 \*   
## MSSubClass90 -4.927e+04 5.728e+03 -8.600 < 2e-16 \*\*\*  
## MSSubClass120 1.710e+04 4.451e+03 3.842 0.000127 \*\*\*  
## MSSubClass160 2.592e+03 6.575e+03 0.394 0.693450   
## MSSubClass180 -1.113e+04 1.230e+04 -0.905 0.365682   
## MSSubClass190 -4.917e+04 7.718e+03 -6.371 2.53e-10 \*\*\*  
## YearRebuilt 4.619e+03 2.365e+03 1.953 0.051011 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 37130 on 1431 degrees of freedom  
## Multiple R-squared: 0.7695, Adjusted R-squared: 0.7656   
## F-statistic: 199 on 24 and 1431 DF, p-value: < 2.2e-16

* Reduce the adjusted R-square

# Recode total area

newtotalArea = train$`1stFlrSF` + train$`2ndFlrSF`  
model.area = lm(Price ~ newtotalArea + FullBath+GarageArea+BsmtExposure\_new+Fireplace + OverallCond + MSSubClass + YearRebuilt + copy\_2$BsmtFinSF1 + copy\_2$BsmtUnfSF, data = copy\_2); summary(model.area)

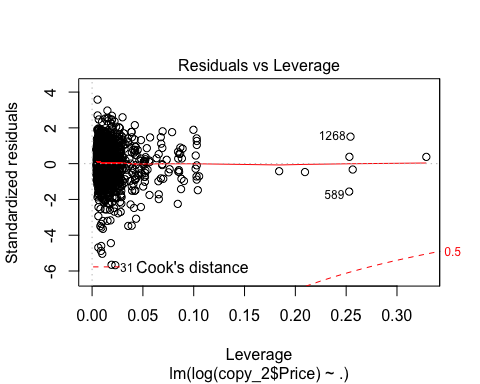
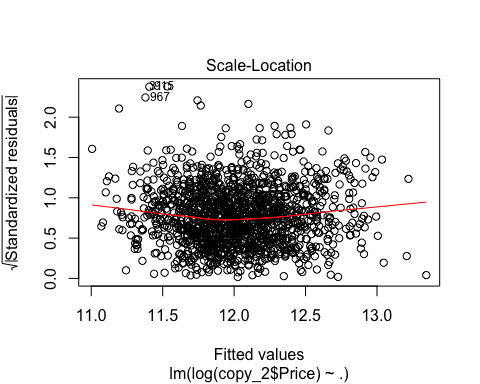
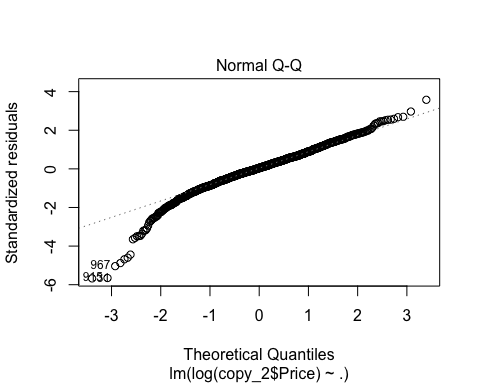
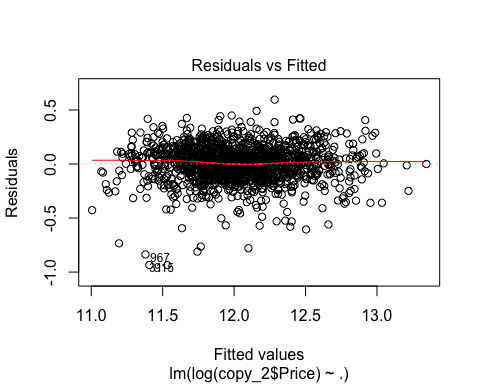
##   
## Call:  
## lm(formula = Price ~ newtotalArea + FullBath + GarageArea + BsmtExposure\_new +   
## Fireplace + OverallCond + MSSubClass + YearRebuilt + copy\_2$BsmtFinSF1 +   
## copy\_2$BsmtUnfSF, data = copy\_2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -134165 -18383 -932 16508 234256   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -51514.070 6725.475 -7.660 3.42e-14 \*\*\*  
## newtotalArea 77.042 3.313 23.253 < 2e-16 \*\*\*  
## FullBath 5124.068 2563.650 1.999 0.045825 \*   
## GarageArea 65.886 5.601 11.763 < 2e-16 \*\*\*  
## BsmtExposure\_new 20633.174 2242.678 9.200 < 2e-16 \*\*\*  
## Fireplace 7593.287 2219.017 3.422 0.000639 \*\*\*  
## OverallCond 4693.171 919.387 5.105 3.76e-07 \*\*\*  
## MSSubClass30 -13391.819 4954.167 -2.703 0.006950 \*\*   
## MSSubClass40 -10909.832 17754.933 -0.614 0.539004   
## MSSubClass45 1362.336 10466.297 0.130 0.896455   
## MSSubClass50 -24062.202 3889.393 -6.187 8.01e-10 \*\*\*  
## MSSubClass60 9547.655 3499.490 2.728 0.006444 \*\*   
## MSSubClass70 -25227.636 5543.493 -4.551 5.79e-06 \*\*\*  
## MSSubClass75 -33414.046 9530.383 -3.506 0.000469 \*\*\*  
## MSSubClass80 -13098.220 5135.024 -2.551 0.010852 \*   
## MSSubClass85 -16348.554 8267.033 -1.978 0.048170 \*   
## MSSubClass90 -41200.263 5476.863 -7.523 9.44e-14 \*\*\*  
## MSSubClass120 11368.959 4158.821 2.734 0.006340 \*\*   
## MSSubClass160 -2708.186 5055.662 -0.536 0.592267   
## MSSubClass180 -4978.077 11626.268 -0.428 0.668588   
## MSSubClass190 -48492.230 6909.606 -7.018 3.46e-12 \*\*\*  
## YearRebuilt 5559.312 2243.997 2.477 0.013348 \*   
## copy\_2$BsmtFinSF1 56.866 3.805 14.944 < 2e-16 \*\*\*  
## copy\_2$BsmtUnfSF 41.990 3.346 12.551 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 35250 on 1432 degrees of freedom  
## Multiple R-squared: 0.7921, Adjusted R-squared: 0.7888   
## F-statistic: 237.3 on 23 and 1432 DF, p-value: < 2.2e-16

* Reduce the adjusted R-square

# Model without Year Built

copy\_2$YearRebuilt = NULL  
model1 = lm(log(copy\_2$Price) ~., data = copy\_2); summary(model1); plot(model1);

##   
## Call:  
## lm(formula = log(copy\_2$Price) ~ ., data = copy\_2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.93599 -0.08883 0.00719 0.10107 0.59542   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.073e+01 3.368e-02 318.644 < 2e-16 \*\*\*  
## MSSubClass30 -2.279e-01 2.246e-02 -10.145 < 2e-16 \*\*\*  
## MSSubClass40 -8.792e-02 8.417e-02 -1.045 0.296426   
## MSSubClass45 -8.901e-02 4.937e-02 -1.803 0.071617 .   
## MSSubClass50 -1.233e-01 2.139e-02 -5.765 9.97e-09 \*\*\*  
## MSSubClass60 1.110e-01 2.648e-02 4.192 2.94e-05 \*\*\*  
## MSSubClass70 -9.602e-02 3.102e-02 -3.095 0.002004 \*\*   
## MSSubClass75 -1.532e-01 4.878e-02 -3.141 0.001718 \*\*   
## MSSubClass80 -9.930e-03 2.465e-02 -0.403 0.687102   
## MSSubClass85 -4.167e-02 3.945e-02 -1.056 0.291104   
## MSSubClass90 -1.793e-01 2.600e-02 -6.897 7.95e-12 \*\*\*  
## MSSubClass120 1.175e-01 2.002e-02 5.866 5.54e-09 \*\*\*  
## MSSubClass160 1.597e-02 2.950e-02 0.541 0.588407   
## MSSubClass180 -1.321e-01 5.536e-02 -2.387 0.017133 \*   
## MSSubClass190 -2.512e-01 3.451e-02 -7.278 5.57e-13 \*\*\*  
## LotArea 1.734e-06 4.827e-07 3.593 0.000338 \*\*\*  
## OverallCond 4.829e-02 4.251e-03 11.358 < 2e-16 \*\*\*  
## BsmtFinSF1 2.536e-04 1.901e-05 13.338 < 2e-16 \*\*\*  
## BsmtUnfSF 2.235e-04 1.700e-05 13.144 < 2e-16 \*\*\*  
## `1stFlrSF` 3.604e-04 2.156e-05 16.720 < 2e-16 \*\*\*  
## `2ndFlrSF` 3.269e-04 2.414e-05 13.543 < 2e-16 \*\*\*  
## FullBath 4.579e-02 1.214e-02 3.773 0.000168 \*\*\*  
## GarageArea 3.657e-04 2.659e-05 13.757 < 2e-16 \*\*\*  
## BsmtExposure\_new 7.375e-02 1.073e-02 6.873 9.38e-12 \*\*\*  
## Fireplace 6.389e-02 1.065e-02 5.999 2.52e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1669 on 1431 degrees of freedom  
## Multiple R-squared: 0.8254, Adjusted R-squared: 0.8225   
## F-statistic: 281.8 on 24 and 1431 DF, p-value: < 2.2e-16



#Confident interval  
confint(model1)

## 2.5 % 97.5 %  
## (Intercept) 1.066593e+01 1.079806e+01  
## MSSubClass30 -2.719352e-01 -1.838091e-01  
## MSSubClass40 -2.530266e-01 7.719365e-02  
## MSSubClass45 -1.858610e-01 7.837247e-03  
## MSSubClass50 -1.652792e-01 -8.136023e-02  
## MSSubClass60 5.904611e-02 1.629188e-01  
## MSSubClass70 -1.568680e-01 -3.516894e-02  
## MSSubClass75 -2.489119e-01 -5.753396e-02  
## MSSubClass80 -5.828225e-02 3.842150e-02  
## MSSubClass85 -1.190569e-01 3.572559e-02  
## MSSubClass90 -2.302886e-01 -1.283003e-01  
## MSSubClass120 7.818210e-02 1.567430e-01  
## MSSubClass160 -4.190370e-02 7.384098e-02  
## MSSubClass180 -2.407066e-01 -2.352708e-02  
## MSSubClass190 -3.188999e-01 -1.834918e-01  
## LotArea 7.875276e-07 2.681233e-06  
## OverallCond 3.994666e-02 5.662627e-02  
## BsmtFinSF1 2.163225e-04 2.909220e-04  
## BsmtUnfSF 1.901044e-04 2.567984e-04  
## `1stFlrSF` 3.181567e-04 4.027320e-04  
## `2ndFlrSF` 2.795655e-04 3.742731e-04  
## FullBath 2.198108e-02 6.959260e-02  
## GarageArea 3.135837e-04 4.178868e-04  
## BsmtExposure\_new 5.270106e-02 9.480255e-02  
## Fireplace 4.299503e-02 8.477908e-02

#Vif test to see the significant of each factors in the model   
car::vif(model1)

## GVIF Df GVIF^(1/(2\*Df))  
## MSSubClass 13.234272 14 1.096631  
## LotArea 1.183463 1 1.087871  
## OverallCond 1.171721 1 1.082461  
## BsmtFinSF1 3.496494 1 1.869891  
## BsmtUnfSF 2.951986 1 1.718135  
## `1stFlrSF` 3.311099 1 1.819643  
## `2ndFlrSF` 5.668884 1 2.380942  
## FullBath 1.866161 1 1.366075  
## GarageArea 1.659280 1 1.288130  
## BsmtExposure\_new 1.308598 1 1.143940  
## Fireplace 1.478371 1 1.215883

* All VIF’s values are smaller than 5 and greater than 1. They also almost close to 1.
* No multicolinearity

anova(model1)

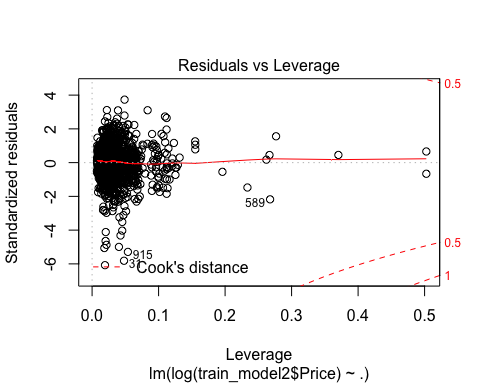
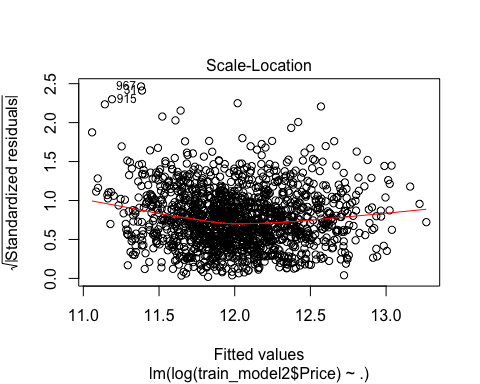
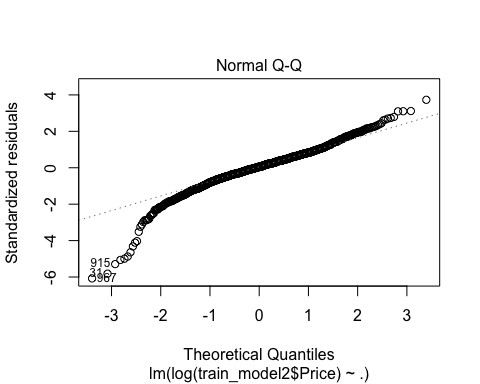
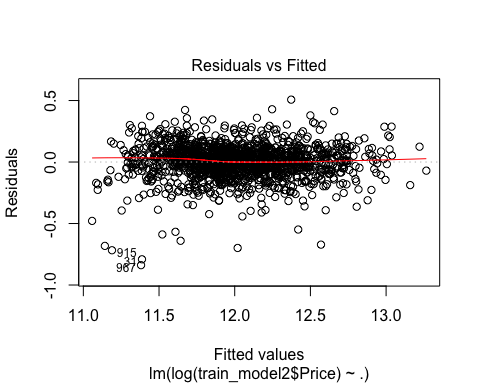
## Analysis of Variance Table  
##   
## Response: log(copy\_2$Price)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## MSSubClass 14 75.318 5.380 193.157 < 2.2e-16 \*\*\*  
## LotArea 1 10.283 10.283 369.187 < 2.2e-16 \*\*\*  
## OverallCond 1 0.586 0.586 21.041 4.888e-06 \*\*\*  
## BsmtFinSF1 1 17.891 17.891 642.337 < 2.2e-16 \*\*\*  
## BsmtUnfSF 1 49.561 49.561 1779.407 < 2.2e-16 \*\*\*  
## `1stFlrSF` 1 19.636 19.636 705.015 < 2.2e-16 \*\*\*  
## `2ndFlrSF` 1 6.249 6.249 224.368 < 2.2e-16 \*\*\*  
## FullBath 1 0.726 0.726 26.054 3.765e-07 \*\*\*  
## GarageArea 1 5.844 5.844 209.824 < 2.2e-16 \*\*\*  
## BsmtExposure\_new 1 1.306 1.306 46.900 1.106e-11 \*\*\*  
## Fireplace 1 1.002 1.002 35.983 2.517e-09 \*\*\*  
## Residuals 1431 39.857 0.028   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* MSE model 1: 0.028
* All explanatory is significant

### Model with Neighboor indicator

train\_model2 = copy\_2  
train\_model2$Neighboor = train$Neighborhood  
train\_model2$MSSubClass = as.factor(copy\_2$MSSubClass)  
  
model2 = lm(log(train\_model2$Price) ~.,data = train\_model2); summary(model2); plot(model2);

##   
## Call:  
## lm(formula = log(train\_model2$Price) ~ ., data = train\_model2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.83856 -0.06726 0.00735 0.08032 0.50724   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.095e+01 5.056e-02 216.595 < 2e-16 \*\*\*  
## MSSubClass30 -1.601e-01 2.246e-02 -7.127 1.64e-12 \*\*\*  
## MSSubClass40 -3.601e-02 7.106e-02 -0.507 0.612422   
## MSSubClass45 -5.109e-02 4.278e-02 -1.194 0.232544   
## MSSubClass50 -5.859e-02 2.027e-02 -2.890 0.003907 \*\*   
## MSSubClass60 5.986e-02 2.288e-02 2.617 0.008976 \*\*   
## MSSubClass70 -5.103e-02 2.875e-02 -1.775 0.076171 .   
## MSSubClass75 -2.340e-02 4.319e-02 -0.542 0.588059   
## MSSubClass80 1.188e-02 2.085e-02 0.570 0.568950   
## MSSubClass85 -6.455e-03 3.325e-02 -0.194 0.846074   
## MSSubClass90 -1.191e-01 2.217e-02 -5.374 9.02e-08 \*\*\*  
## MSSubClass120 9.044e-03 1.990e-02 0.455 0.649495   
## MSSubClass160 -4.971e-02 3.114e-02 -1.597 0.110582   
## MSSubClass180 2.044e-02 5.583e-02 0.366 0.714380   
## MSSubClass190 -1.613e-01 3.098e-02 -5.208 2.19e-07 \*\*\*  
## LotArea 1.422e-06 4.316e-07 3.294 0.001011 \*\*   
## OverallCond 5.911e-02 3.710e-03 15.930 < 2e-16 \*\*\*  
## BsmtFinSF1 2.032e-04 1.630e-05 12.467 < 2e-16 \*\*\*  
## BsmtUnfSF 1.419e-04 1.495e-05 9.492 < 2e-16 \*\*\*  
## `1stFlrSF` 3.572e-04 1.865e-05 19.149 < 2e-16 \*\*\*  
## `2ndFlrSF` 2.998e-04 2.072e-05 14.471 < 2e-16 \*\*\*  
## FullBath 3.910e-02 1.027e-02 3.808 0.000146 \*\*\*  
## GarageArea 2.398e-04 2.373e-05 10.108 < 2e-16 \*\*\*  
## BsmtExposure\_new 3.768e-02 9.418e-03 4.000 6.66e-05 \*\*\*  
## Fireplace 5.528e-02 9.377e-03 5.895 4.69e-09 \*\*\*  
## NeighboorBlueste -1.288e-01 1.086e-01 -1.186 0.235865   
## NeighboorBrDale -2.285e-01 5.710e-02 -4.001 6.63e-05 \*\*\*  
## NeighboorBrkSide -1.971e-01 4.569e-02 -4.314 1.71e-05 \*\*\*  
## NeighboorClearCr -9.203e-02 4.782e-02 -1.925 0.054488 .   
## NeighboorCollgCr -2.579e-02 4.013e-02 -0.643 0.520612   
## NeighboorCrawfor -4.618e-02 4.496e-02 -1.027 0.304580   
## NeighboorEdwards -2.203e-01 4.255e-02 -5.179 2.56e-07 \*\*\*  
## NeighboorGilbert -1.654e-02 4.237e-02 -0.390 0.696249   
## NeighboorIDOTRR -3.454e-01 4.767e-02 -7.244 7.14e-13 \*\*\*  
## NeighboorMeadowV -3.773e-01 5.712e-02 -6.605 5.63e-11 \*\*\*  
## NeighboorMitchel -1.517e-01 4.388e-02 -3.458 0.000561 \*\*\*  
## NeighboorNAmes -2.045e-01 4.068e-02 -5.027 5.63e-07 \*\*\*  
## NeighboorNoRidge -1.358e-02 4.557e-02 -0.298 0.765697   
## NeighboorNPkVill -1.273e-01 6.007e-02 -2.119 0.034296 \*   
## NeighboorNridgHt 1.332e-01 4.012e-02 3.321 0.000921 \*\*\*  
## NeighboorNWAmes -1.811e-01 4.235e-02 -4.277 2.02e-05 \*\*\*  
## NeighboorOldTown -2.842e-01 4.358e-02 -6.522 9.67e-11 \*\*\*  
## NeighboorSawyer -2.126e-01 4.310e-02 -4.933 9.06e-07 \*\*\*  
## NeighboorSawyerW -8.679e-02 4.258e-02 -2.038 0.041717 \*   
## NeighboorSomerst 1.056e-01 4.121e-02 2.562 0.010509 \*   
## NeighboorStoneBr 1.482e-01 4.551e-02 3.257 0.001154 \*\*   
## NeighboorSWISU -2.064e-01 5.056e-02 -4.082 4.72e-05 \*\*\*  
## NeighboorTimber -1.315e-02 4.532e-02 -0.290 0.771763   
## NeighboorVeenker 7.784e-03 5.640e-02 0.138 0.890244   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1394 on 1407 degrees of freedom  
## Multiple R-squared: 0.8803, Adjusted R-squared: 0.8762   
## F-statistic: 215.5 on 48 and 1407 DF, p-value: < 2.2e-16



anova(model2)

## Analysis of Variance Table  
##   
## Response: log(train\_model2$Price)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## MSSubClass 14 75.318 5.380 276.973 < 2.2e-16 \*\*\*  
## LotArea 1 10.283 10.283 529.388 < 2.2e-16 \*\*\*  
## OverallCond 1 0.586 0.586 30.172 4.687e-08 \*\*\*  
## BsmtFinSF1 1 17.891 17.891 921.065 < 2.2e-16 \*\*\*  
## BsmtUnfSF 1 49.561 49.561 2551.542 < 2.2e-16 \*\*\*  
## `1stFlrSF` 1 19.636 19.636 1010.941 < 2.2e-16 \*\*\*  
## `2ndFlrSF` 1 6.249 6.249 321.727 < 2.2e-16 \*\*\*  
## FullBath 1 0.726 0.726 37.360 1.269e-09 \*\*\*  
## GarageArea 1 5.844 5.844 300.873 < 2.2e-16 \*\*\*  
## BsmtExposure\_new 1 1.306 1.306 67.251 5.331e-16 \*\*\*  
## Fireplace 1 1.002 1.002 51.597 1.101e-12 \*\*\*  
## Neighboor 24 12.527 0.522 26.873 < 2.2e-16 \*\*\*  
## Residuals 1407 27.329 0.019   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* Adjusted R-square: 0.8762
* MSE: 0.019

#Checking outliner  
train\_model2[967,]

## # A tibble: 1 x 13  
## MSSubClass LotArea OverallCond BsmtFinSF1 BsmtUnfSF `1stFlrSF` `2ndFlrSF`  
## <fct> <int> <int> <int> <int> <int> <int>  
## 1 50 5925 6 0 600 600 368  
## # ... with 6 more variables: FullBath <dbl>, GarageArea <int>,  
## # BsmtExposure\_new <dbl>, Fireplace <dbl>, Price <int>, Neighboor <chr>

#Vif test to see the significant of each factors in the model   
car::vif(model2)

## GVIF Df GVIF^(1/(2\*Df))  
## MSSubClass 161.422931 14 1.199101  
## LotArea 1.356563 1 1.164716  
## OverallCond 1.279723 1 1.131248  
## BsmtFinSF1 3.685393 1 1.919738  
## BsmtUnfSF 3.273932 1 1.809401  
## `1stFlrSF` 3.554623 1 1.885371  
## `2ndFlrSF` 5.988434 1 2.447128  
## FullBath 1.915749 1 1.384106  
## GarageArea 1.894868 1 1.376542  
## BsmtExposure\_new 1.445369 1 1.202235  
## Fireplace 1.643314 1 1.281918  
## Neighboor 63.519462 24 1.090337

* All factors has VIF < 5 and VIF > 1.

confint(model2)

## 2.5 % 97.5 %  
## (Intercept) 1.085200e+01 1.105037e+01  
## MSSubClass30 -2.041277e-01 -1.160091e-01  
## MSSubClass40 -1.754120e-01 1.033917e-01  
## MSSubClass45 -1.350128e-01 3.282514e-02  
## MSSubClass50 -9.835321e-02 -1.882513e-02  
## MSSubClass60 1.498380e-02 1.047412e-01  
## MSSubClass70 -1.074291e-01 5.376687e-03  
## MSSubClass75 -1.081321e-01 6.132961e-02  
## MSSubClass80 -2.902048e-02 5.277751e-02  
## MSSubClass85 -7.167114e-02 5.876088e-02  
## MSSubClass90 -1.625891e-01 -7.562875e-02  
## MSSubClass120 -2.998397e-02 4.807122e-02  
## MSSubClass160 -1.107960e-01 1.136722e-02  
## MSSubClass180 -8.908861e-02 1.299651e-01  
## MSSubClass190 -2.220978e-01 -1.005613e-01  
## LotArea 5.751329e-07 2.268291e-06  
## OverallCond 5.182809e-02 6.638517e-02  
## BsmtFinSF1 1.712647e-04 2.352241e-04  
## BsmtUnfSF 1.125842e-04 1.712395e-04  
## `1stFlrSF` 3.205850e-04 3.937658e-04  
## `2ndFlrSF` 2.591818e-04 3.404714e-04  
## FullBath 1.896092e-02 5.924648e-02  
## GarageArea 1.932748e-04 2.863576e-04  
## BsmtExposure\_new 1.919965e-02 5.615066e-02  
## Fireplace 3.688051e-02 7.366982e-02  
## NeighboorBlueste -3.419722e-01 8.428400e-02  
## NeighboorBrDale -3.404825e-01 -1.164612e-01  
## NeighboorBrkSide -2.867387e-01 -1.074852e-01  
## NeighboorClearCr -1.858261e-01 1.774303e-03  
## NeighboorCollgCr -1.045117e-01 5.293729e-02  
## NeighboorCrawfor -1.343855e-01 4.202468e-02  
## NeighboorEdwards -3.038096e-01 -1.368788e-01  
## NeighboorGilbert -9.966029e-02 6.657169e-02  
## NeighboorIDOTRR -4.388676e-01 -2.518345e-01  
## NeighboorMeadowV -4.893025e-01 -2.652059e-01  
## NeighboorMitchel -2.377919e-01 -6.564702e-02  
## NeighboorNAmes -2.842640e-01 -1.246769e-01  
## NeighboorNoRidge -1.029653e-01 7.580227e-02  
## NeighboorNPkVill -2.451115e-01 -9.430272e-03  
## NeighboorNridgHt 5.451320e-02 2.119018e-01  
## NeighboorNWAmes -2.641890e-01 -9.804064e-02  
## NeighboorOldTown -3.696634e-01 -1.987050e-01  
## NeighboorSawyer -2.971495e-01 -1.280637e-01  
## NeighboorSawyerW -1.703271e-01 -3.259479e-03  
## NeighboorSomerst 2.474107e-02 1.864077e-01  
## NeighboorStoneBr 5.893818e-02 2.374962e-01  
## NeighboorSWISU -3.055982e-01 -1.072172e-01  
## NeighboorTimber -1.020428e-01 7.574780e-02  
## NeighboorVeenker -1.028507e-01 1.184191e-01

### Citation

De Cock, Dean. “Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project.” Journal of Statistics Education 19, no. 3 (November 2011). <https://doi.org/10.1080/10691898.2011.11889627>.

Kuhn, Max, and Kjell Johnson. “Data Pre-Processing.” In Applied Predictive Modeling, edited by Max Kuhn and Kjell Johnson, 27–59. New York, NY: Springer New York, 2013. <https://doi.org/10.1007/978-1-4614-6849-3_3>.

“Information About Factors That Determine Property Prices - HomeGuru.” Accessed October 18, 2018. <http://www.homeguru.com.au/house-prices/>.