house\_price\_prediction

Himalya

June 26, 2019

# Load Packages

library(ggplot2)  
library(ggthemes)  
library(scales)  
library(dplyr)  
library(mice)  
library(randomForest)  
library(data.table)  
library(gridExtra)  
library(corrplot)  
library(GGally)  
library(e1071)

# Data Overview

# Reading input train data and test data  
train\_data <- read.csv('train\_house.csv', stringsAsFactors = F)  
test\_data <- read.csv('test\_house.csv', stringsAsFactors = F)  
  
# Dimension of data  
cat('Housing train data has', dim(train\_data)[1], 'rows and', dim(train\_data)[2], 'columns')

## Housing train data has 1460 rows and 81 columns

cat('Housing test data has', dim(test\_data)[1], 'rows and', dim(test\_data)[2], 'columns')

## Housing test data has 1459 rows and 80 columns

# Structure of data  
str(train\_data)

## '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 ...

str(test\_data)

## 'data.frame': 1459 obs. of 80 variables:  
## $ Id : int 1461 1462 1463 1464 1465 1466 1467 1468 1469 1470 ...  
## $ MSSubClass : int 20 20 60 60 120 60 20 60 20 20 ...  
## $ MSZoning : chr "RH" "RL" "RL" "RL" ...  
## $ LotFrontage : int 80 81 74 78 43 75 NA 63 85 70 ...  
## $ LotArea : int 11622 14267 13830 9978 5005 10000 7980 8402 10176 8400 ...  
## $ Street : chr "Pave" "Pave" "Pave" "Pave" ...  
## $ Alley : chr NA NA NA NA ...  
## $ LotShape : chr "Reg" "IR1" "IR1" "IR1" ...  
## $ LandContour : chr "Lvl" "Lvl" "Lvl" "Lvl" ...  
## $ Utilities : chr "AllPub" "AllPub" "AllPub" "AllPub" ...  
## $ LotConfig : chr "Inside" "Corner" "Inside" "Inside" ...  
## $ LandSlope : chr "Gtl" "Gtl" "Gtl" "Gtl" ...  
## $ Neighborhood : chr "NAmes" "NAmes" "Gilbert" "Gilbert" ...  
## $ Condition1 : chr "Feedr" "Norm" "Norm" "Norm" ...  
## $ Condition2 : chr "Norm" "Norm" "Norm" "Norm" ...  
## $ BldgType : chr "1Fam" "1Fam" "1Fam" "1Fam" ...  
## $ HouseStyle : chr "1Story" "1Story" "2Story" "2Story" ...  
## $ OverallQual : int 5 6 5 6 8 6 6 6 7 4 ...  
## $ OverallCond : int 6 6 5 6 5 5 7 5 5 5 ...  
## $ YearBuilt : int 1961 1958 1997 1998 1992 1993 1992 1998 1990 1970 ...  
## $ YearRemodAdd : int 1961 1958 1998 1998 1992 1994 2007 1998 1990 1970 ...  
## $ RoofStyle : chr "Gable" "Hip" "Gable" "Gable" ...  
## $ RoofMatl : chr "CompShg" "CompShg" "CompShg" "CompShg" ...  
## $ Exterior1st : chr "VinylSd" "Wd Sdng" "VinylSd" "VinylSd" ...  
## $ Exterior2nd : chr "VinylSd" "Wd Sdng" "VinylSd" "VinylSd" ...  
## $ MasVnrType : chr "None" "BrkFace" "None" "BrkFace" ...  
## $ MasVnrArea : int 0 108 0 20 0 0 0 0 0 0 ...  
## $ ExterQual : chr "TA" "TA" "TA" "TA" ...  
## $ ExterCond : chr "TA" "TA" "TA" "TA" ...  
## $ Foundation : chr "CBlock" "CBlock" "PConc" "PConc" ...  
## $ BsmtQual : chr "TA" "TA" "Gd" "TA" ...  
## $ BsmtCond : chr "TA" "TA" "TA" "TA" ...  
## $ BsmtExposure : chr "No" "No" "No" "No" ...  
## $ BsmtFinType1 : chr "Rec" "ALQ" "GLQ" "GLQ" ...  
## $ BsmtFinSF1 : int 468 923 791 602 263 0 935 0 637 804 ...  
## $ BsmtFinType2 : chr "LwQ" "Unf" "Unf" "Unf" ...  
## $ BsmtFinSF2 : int 144 0 0 0 0 0 0 0 0 78 ...  
## $ BsmtUnfSF : int 270 406 137 324 1017 763 233 789 663 0 ...  
## $ TotalBsmtSF : int 882 1329 928 926 1280 763 1168 789 1300 882 ...  
## $ Heating : chr "GasA" "GasA" "GasA" "GasA" ...  
## $ HeatingQC : chr "TA" "TA" "Gd" "Ex" ...  
## $ CentralAir : chr "Y" "Y" "Y" "Y" ...  
## $ Electrical : chr "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...  
## $ X1stFlrSF : int 896 1329 928 926 1280 763 1187 789 1341 882 ...  
## $ X2ndFlrSF : int 0 0 701 678 0 892 0 676 0 0 ...  
## $ LowQualFinSF : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ GrLivArea : int 896 1329 1629 1604 1280 1655 1187 1465 1341 882 ...  
## $ BsmtFullBath : int 0 0 0 0 0 0 1 0 1 1 ...  
## $ BsmtHalfBath : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ FullBath : int 1 1 2 2 2 2 2 2 1 1 ...  
## $ HalfBath : int 0 1 1 1 0 1 0 1 1 0 ...  
## $ BedroomAbvGr : int 2 3 3 3 2 3 3 3 2 2 ...  
## $ KitchenAbvGr : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ KitchenQual : chr "TA" "Gd" "TA" "Gd" ...  
## $ TotRmsAbvGrd : int 5 6 6 7 5 7 6 7 5 4 ...  
## $ Functional : chr "Typ" "Typ" "Typ" "Typ" ...  
## $ Fireplaces : int 0 0 1 1 0 1 0 1 1 0 ...  
## $ FireplaceQu : chr NA NA "TA" "Gd" ...  
## $ GarageType : chr "Attchd" "Attchd" "Attchd" "Attchd" ...  
## $ GarageYrBlt : int 1961 1958 1997 1998 1992 1993 1992 1998 1990 1970 ...  
## $ GarageFinish : chr "Unf" "Unf" "Fin" "Fin" ...  
## $ GarageCars : int 1 1 2 2 2 2 2 2 2 2 ...  
## $ GarageArea : int 730 312 482 470 506 440 420 393 506 525 ...  
## $ GarageQual : chr "TA" "TA" "TA" "TA" ...  
## $ GarageCond : chr "TA" "TA" "TA" "TA" ...  
## $ PavedDrive : chr "Y" "Y" "Y" "Y" ...  
## $ WoodDeckSF : int 140 393 212 360 0 157 483 0 192 240 ...  
## $ OpenPorchSF : int 0 36 34 36 82 84 21 75 0 0 ...  
## $ EnclosedPorch: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ X3SsnPorch : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ScreenPorch : int 120 0 0 0 144 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 "MnPrv" NA "MnPrv" NA ...  
## $ MiscFeature : chr NA "Gar2" NA NA ...  
## $ MiscVal : int 0 12500 0 0 0 0 500 0 0 0 ...  
## $ MoSold : int 6 6 3 6 1 4 3 5 2 4 ...  
## $ YrSold : int 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...  
## $ SaleType : chr "WD" "WD" "WD" "WD" ...  
## $ SaleCondition: chr "Normal" "Normal" "Normal" "Normal" ...

# Summary of numerical data  
summary(train\_data[,sapply(train\_data[,1:81], typeof) == "integer"])

## Id MSSubClass LotFrontage LotArea   
## Min. : 1.0 Min. : 20.0 Min. : 21.00 Min. : 1300   
## 1st Qu.: 365.8 1st Qu.: 20.0 1st Qu.: 59.00 1st Qu.: 7554   
## Median : 730.5 Median : 50.0 Median : 69.00 Median : 9478   
## Mean : 730.5 Mean : 56.9 Mean : 70.05 Mean : 10517   
## 3rd Qu.:1095.2 3rd Qu.: 70.0 3rd Qu.: 80.00 3rd Qu.: 11602   
## Max. :1460.0 Max. :190.0 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 :1973 Median :1994   
## Mean : 6.099 Mean :5.575 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.: 223.0   
## Median : 0.0 Median : 383.5 Median : 0.00 Median : 477.5   
## Mean : 103.7 Mean : 443.6 Mean : 46.55 Mean : 567.2   
## 3rd Qu.: 166.0 3rd Qu.: 712.2 3rd Qu.: 0.00 3rd Qu.: 808.0   
## Max. :1600.0 Max. :5644.0 Max. :1474.00 Max. :2336.0   
## NA's :8   
## TotalBsmtSF X1stFlrSF X2ndFlrSF LowQualFinSF   
## Min. : 0.0 Min. : 334 Min. : 0 Min. : 0.000   
## 1st Qu.: 795.8 1st Qu.: 882 1st Qu.: 0 1st Qu.: 0.000   
## Median : 991.5 Median :1087 Median : 0 Median : 0.000   
## Mean :1057.4 Mean :1163 Mean : 347 Mean : 5.845   
## 3rd Qu.:1298.2 3rd Qu.:1391 3rd Qu.: 728 3rd Qu.: 0.000   
## Max. :6110.0 Max. :4692 Max. :2065 Max. :572.000   
##   
## GrLivArea BsmtFullBath BsmtHalfBath FullBath   
## Min. : 334 Min. :0.0000 Min. :0.00000 Min. :0.000   
## 1st Qu.:1130 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:1.000   
## Median :1464 Median :0.0000 Median :0.00000 Median :2.000   
## Mean :1515 Mean :0.4253 Mean :0.05753 Mean :1.565   
## 3rd Qu.:1777 3rd Qu.:1.0000 3rd Qu.:0.00000 3rd Qu.:2.000   
## Max. :5642 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 :0.3829 Mean :2.866 Mean :1.047 Mean : 6.518   
## 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. :0.000 Min. :1900 Min. :0.000 Min. : 0.0   
## 1st Qu.:0.000 1st Qu.:1961 1st Qu.:1.000 1st Qu.: 334.5   
## Median :1.000 Median :1980 Median :2.000 Median : 480.0   
## Mean :0.613 Mean :1979 Mean :1.767 Mean : 473.0   
## 3rd Qu.:1.000 3rd Qu.:2002 3rd Qu.:2.000 3rd Qu.: 576.0   
## Max. :3.000 Max. :2010 Max. :4.000 Max. :1418.0   
## NA's :81   
## WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch   
## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00   
## Median : 0.00 Median : 25.00 Median : 0.00 Median : 0.00   
## Mean : 94.24 Mean : 46.66 Mean : 21.95 Mean : 3.41   
## 3rd Qu.:168.00 3rd Qu.: 68.00 3rd Qu.: 0.00 3rd Qu.: 0.00   
## Max. :857.00 Max. :547.00 Max. :552.00 Max. :508.00   
##   
## ScreenPorch PoolArea MiscVal MoSold   
## Min. : 0.00 Min. : 0.000 Min. : 0.00 Min. : 1.000   
## 1st Qu.: 0.00 1st Qu.: 0.000 1st Qu.: 0.00 1st Qu.: 5.000   
## Median : 0.00 Median : 0.000 Median : 0.00 Median : 6.000   
## Mean : 15.06 Mean : 2.759 Mean : 43.49 Mean : 6.322   
## 3rd Qu.: 0.00 3rd Qu.: 0.000 3rd Qu.: 0.00 3rd Qu.: 8.000   
## Max. :480.00 Max. :738.000 Max. :15500.00 Max. :12.000   
##   
## YrSold SalePrice   
## Min. :2006 Min. : 34900   
## 1st Qu.:2007 1st Qu.:129975   
## Median :2008 Median :163000   
## Mean :2008 Mean :180921   
## 3rd Qu.:2009 3rd Qu.:214000   
## Max. :2010 Max. :755000   
##

# Percentage of missing data  
cat('Percentage of missing data in train is:', round(sum(is.na(train\_data) / (nrow(train\_data) \* ncol(train\_data))),3))

## Percentage of missing data in train is: 0.059

cat('Percentage of missing data in test is:', round(sum(is.na(test\_data) / (nrow(test\_data) \* ncol(test\_data))),3))

## Percentage of missing data in test is: 0.06

# Combine Data

#### Test Dataset has no “Salesprice” variable. So creating it for prediction

# Combining rows of train and test, with "salesprice" variable in test as NA  
house\_data <- bind\_rows(train\_data,test\_data)  
  
# Data exploration  
str(house\_data)

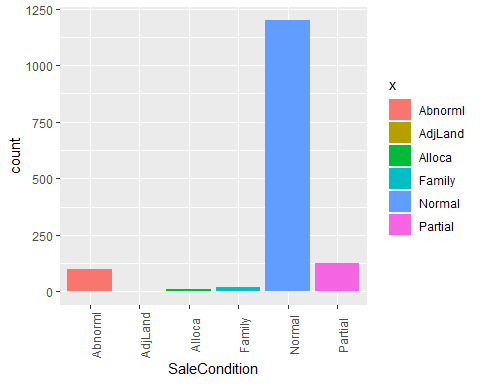
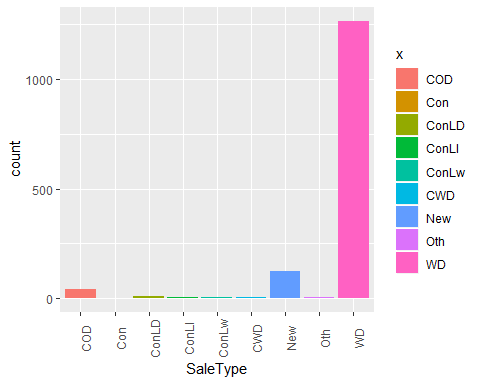
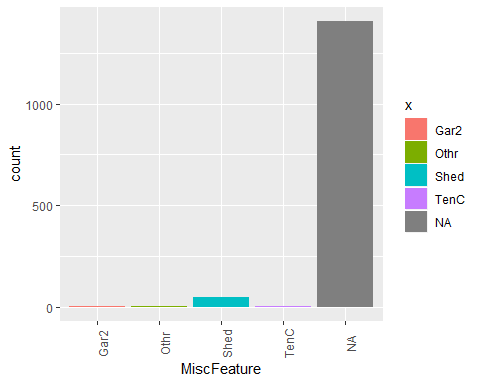
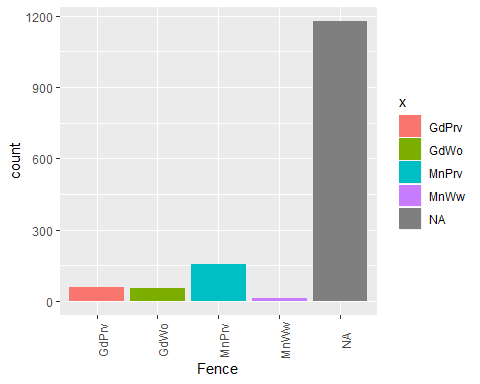
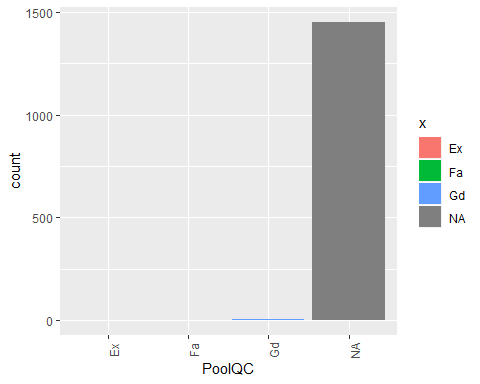
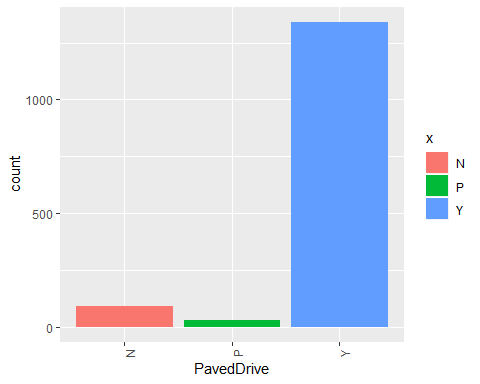
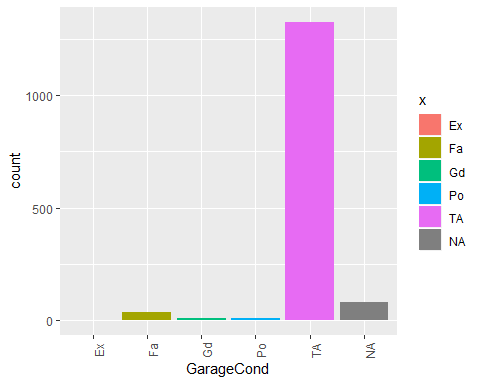
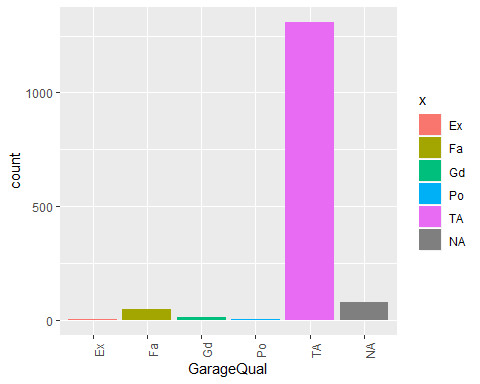
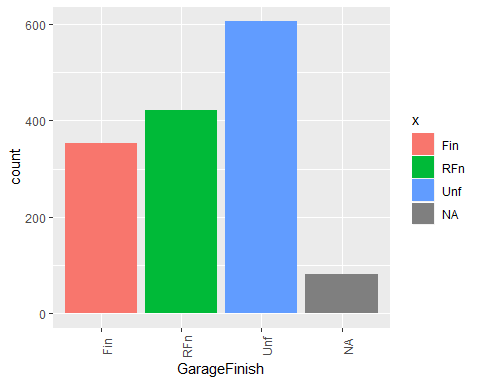
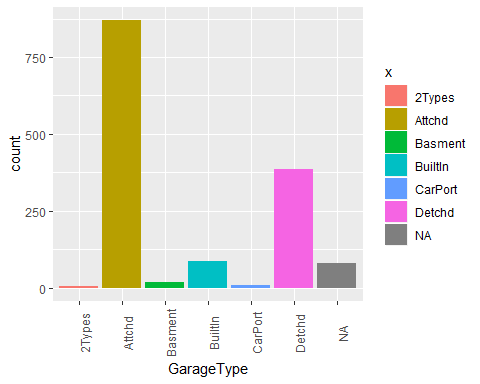
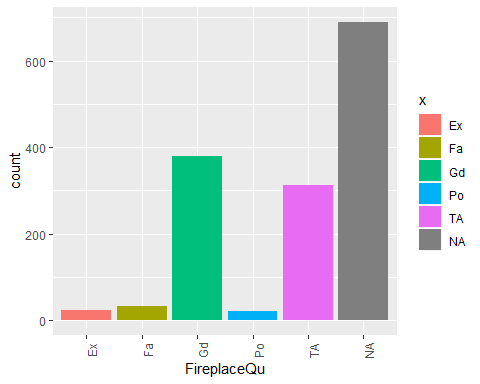
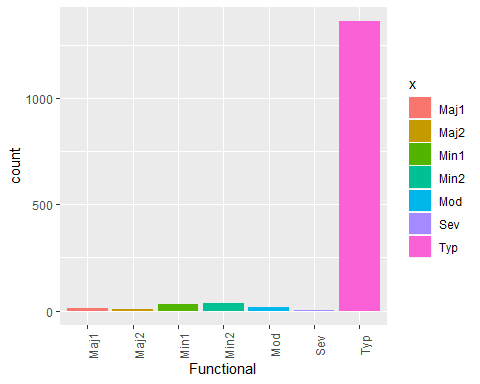
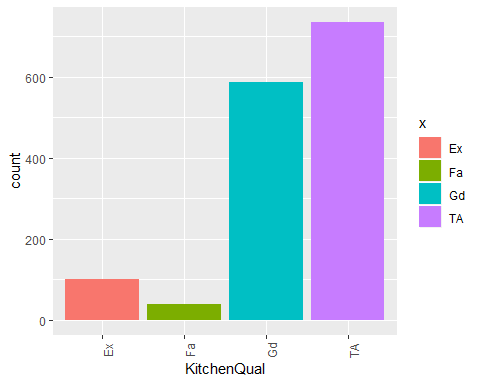
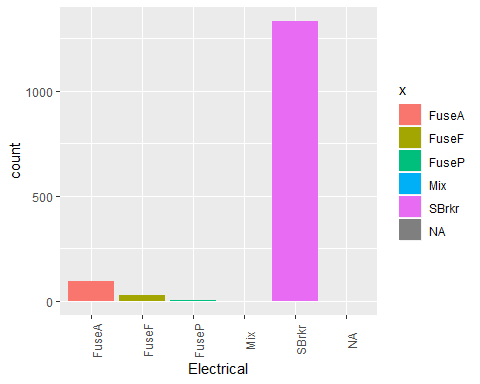
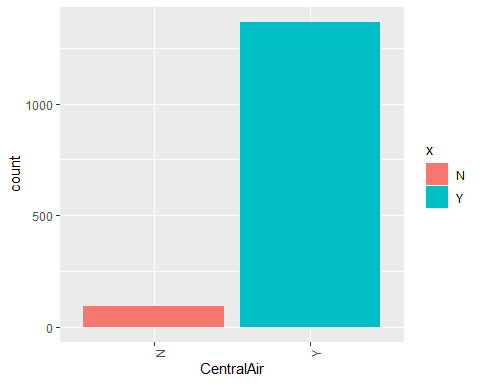
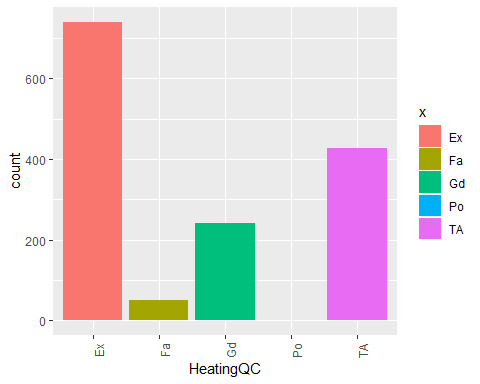
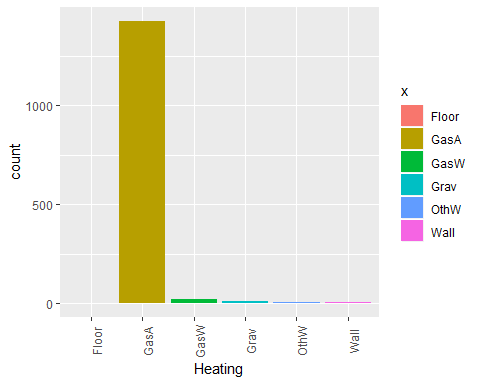
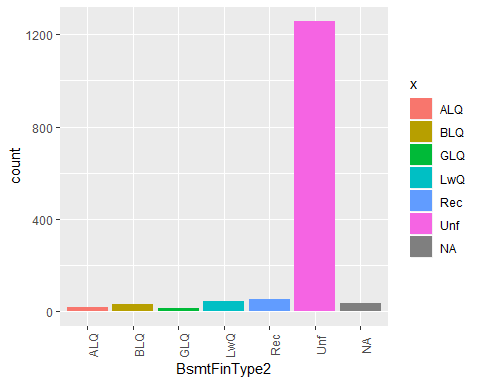
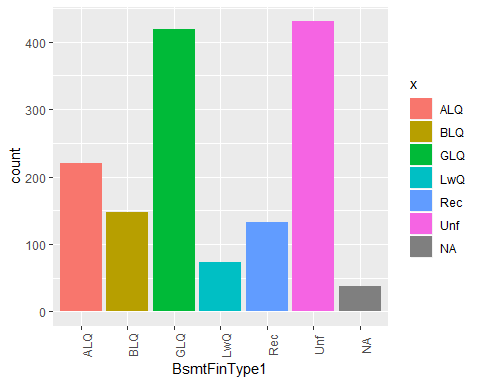
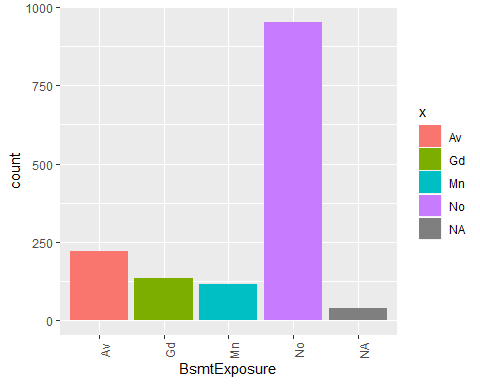
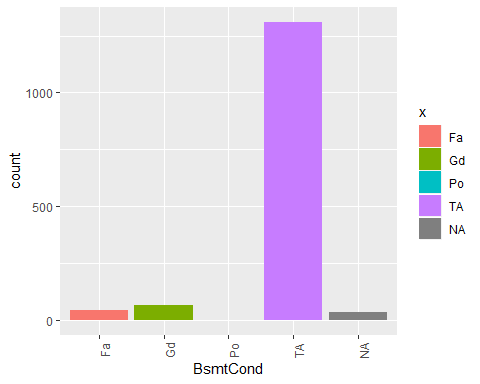
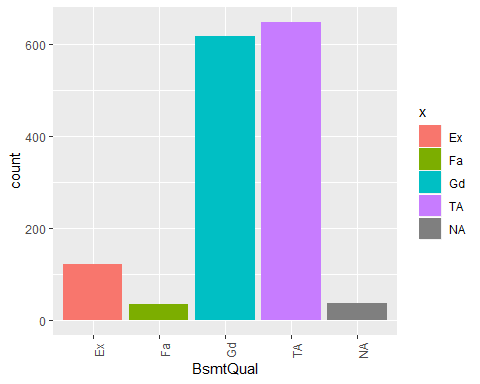
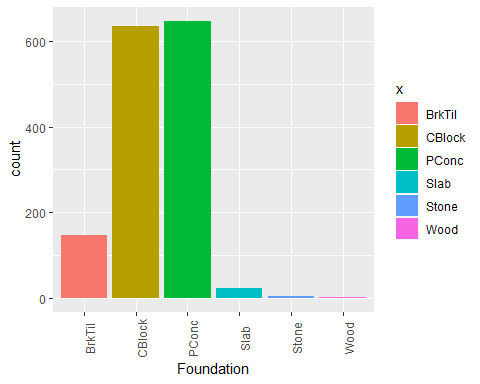
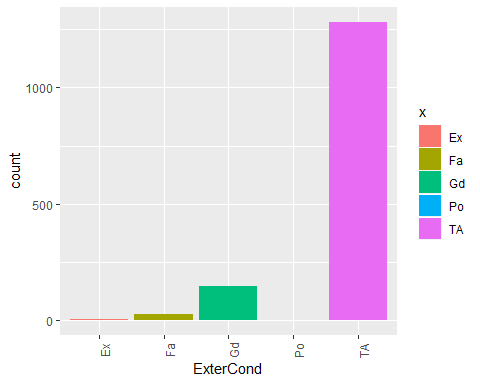
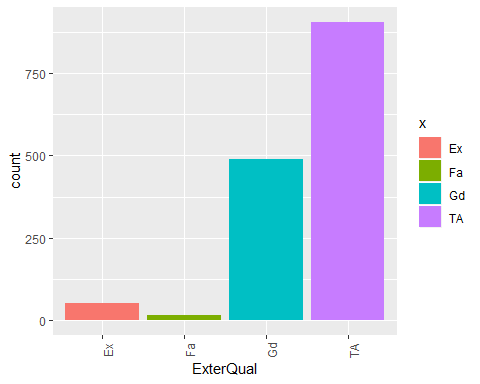
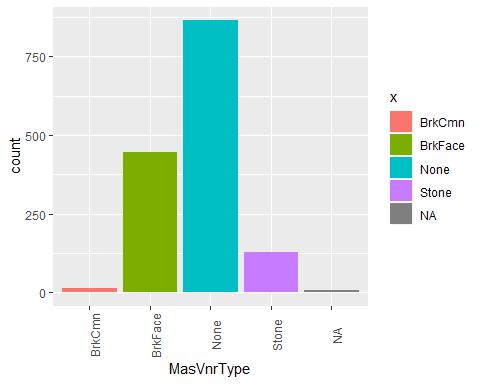
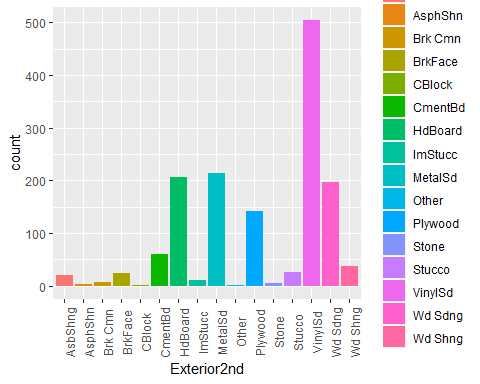
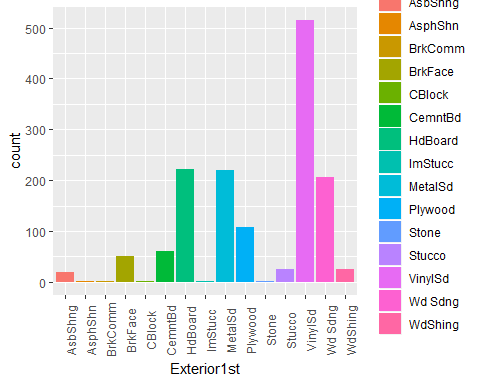
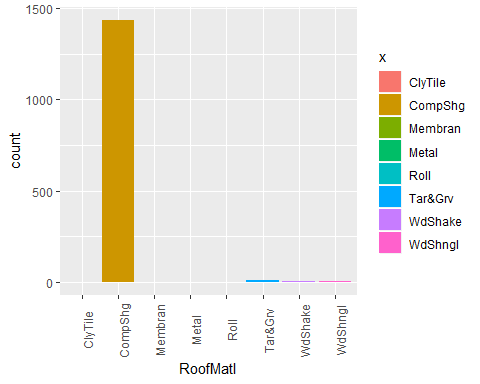
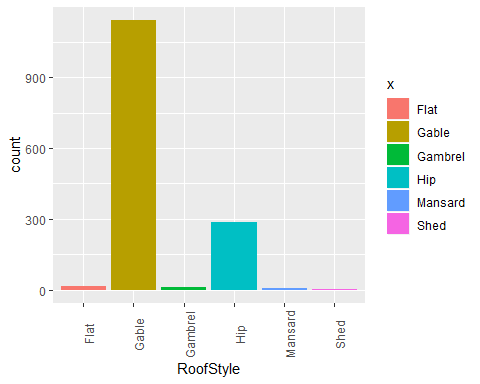
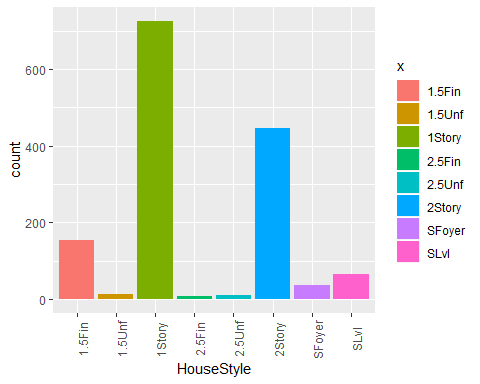
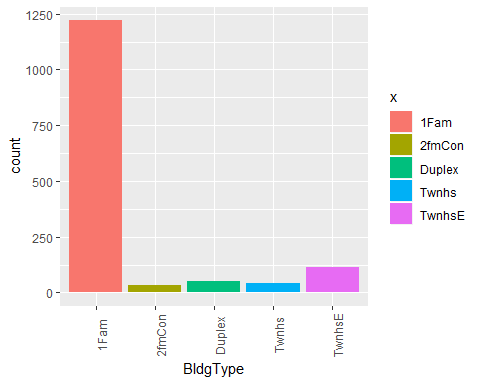
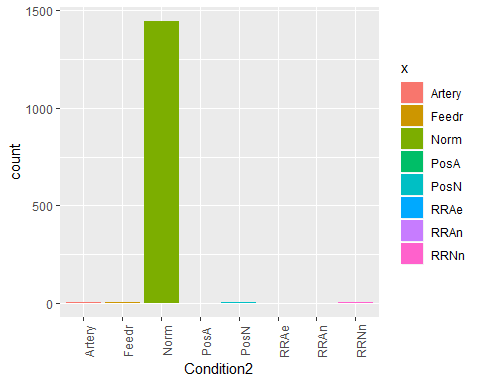
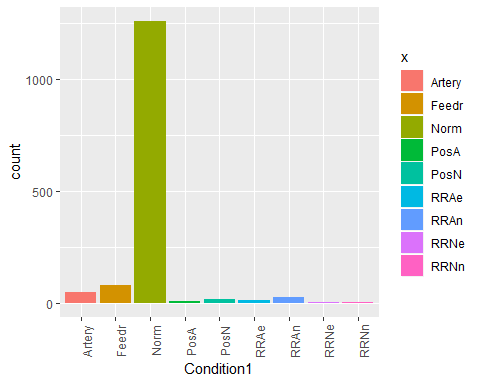
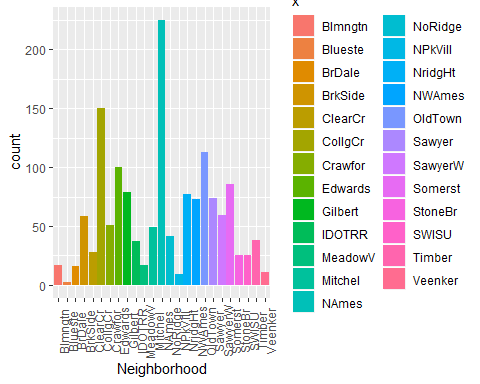
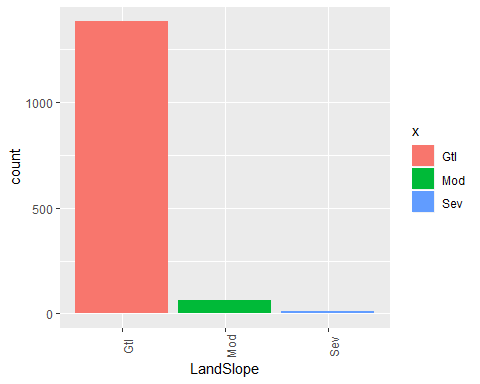
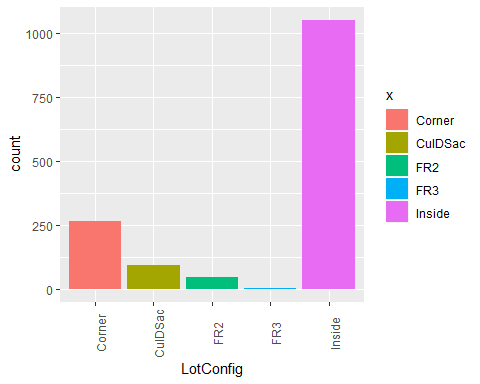
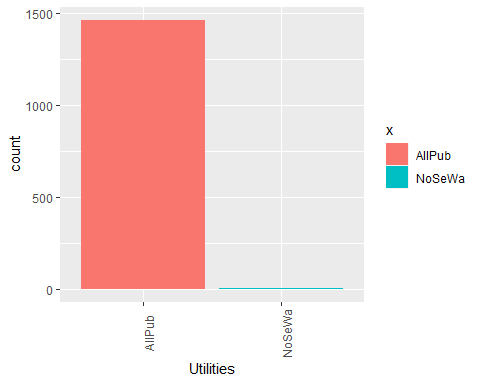
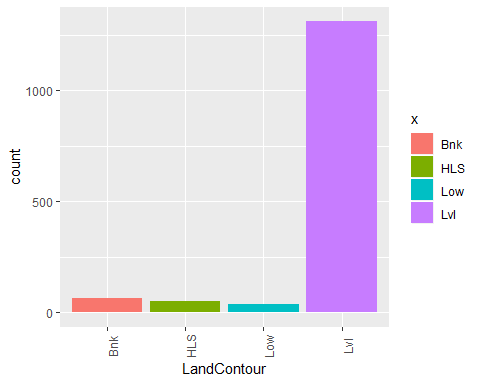
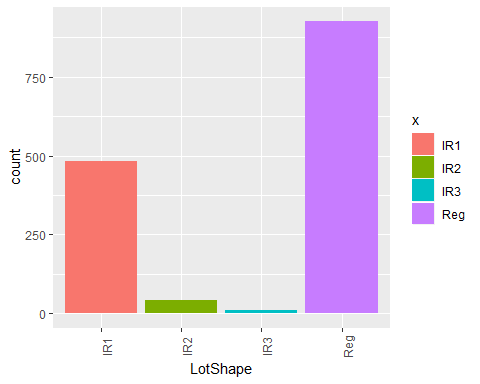
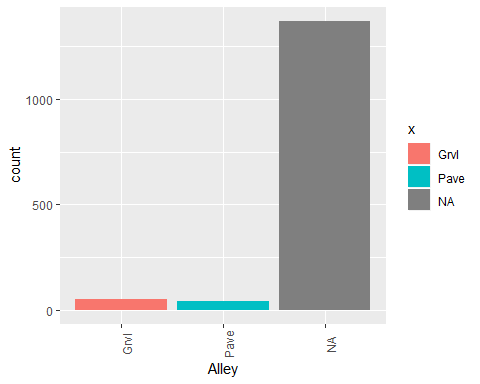
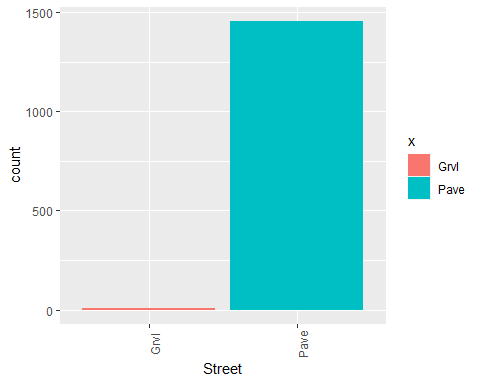
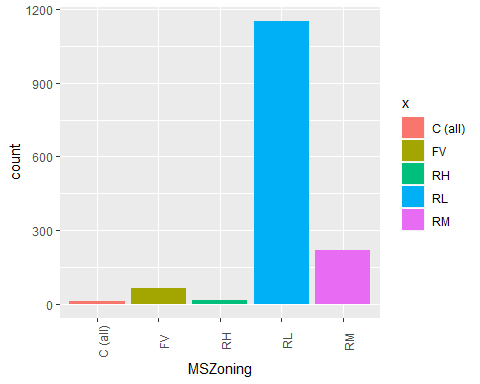
## 'data.frame': 2919 obs. of 81 variables:  
## $ Id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ MSSubClass : int 60 20 60 70 60 50 20 60 50 190 ...  
## $ MSZoning : chr "RL" "RL" "RL" "RL" ...  
## $ LotFrontage : int 65 80 68 60 84 85 75 NA 51 50 ...  
## $ LotArea : int 8450 9600 11250 9550 14260 14115 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(house\_data)

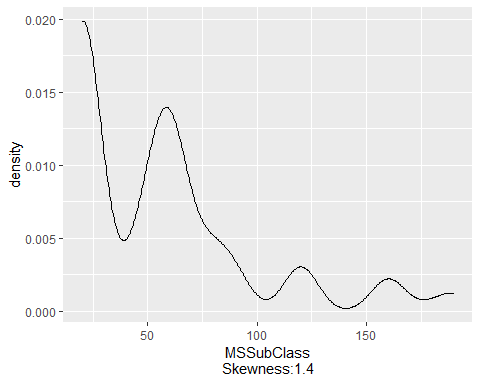
## Id MSSubClass MSZoning LotFrontage   
## Min. : 1.0 Min. : 20.00 Length:2919 Min. : 21.00   
## 1st Qu.: 730.5 1st Qu.: 20.00 Class :character 1st Qu.: 59.00   
## Median :1460.0 Median : 50.00 Mode :character Median : 68.00   
## Mean :1460.0 Mean : 57.14 Mean : 69.31   
## 3rd Qu.:2189.5 3rd Qu.: 70.00 3rd Qu.: 80.00   
## Max. :2919.0 Max. :190.00 Max. :313.00   
## NA's :486   
## LotArea Street Alley LotShape   
## Min. : 1300 Length:2919 Length:2919 Length:2919   
## 1st Qu.: 7478 Class :character Class :character Class :character   
## Median : 9453 Mode :character Mode :character Mode :character   
## Mean : 10168   
## 3rd Qu.: 11570   
## Max. :215245   
##   
## LandContour Utilities LotConfig   
## Length:2919 Length:2919 Length:2919   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## LandSlope Neighborhood Condition1   
## Length:2919 Length:2919 Length:2919   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Condition2 BldgType HouseStyle OverallQual   
## Length:2919 Length:2919 Length:2919 Min. : 1.000   
## Class :character Class :character Class :character 1st Qu.: 5.000   
## Mode :character Mode :character Mode :character Median : 6.000   
## Mean : 6.089   
## 3rd Qu.: 7.000   
## Max. :10.000   
##   
## OverallCond YearBuilt YearRemodAdd RoofStyle   
## Min. :1.000 Min. :1872 Min. :1950 Length:2919   
## 1st Qu.:5.000 1st Qu.:1954 1st Qu.:1965 Class :character   
## Median :5.000 Median :1973 Median :1993 Mode :character   
## Mean :5.565 Mean :1971 Mean :1984   
## 3rd Qu.:6.000 3rd Qu.:2001 3rd Qu.:2004   
## Max. :9.000 Max. :2010 Max. :2010   
##   
## RoofMatl Exterior1st Exterior2nd   
## Length:2919 Length:2919 Length:2919   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## MasVnrType MasVnrArea ExterQual ExterCond   
## Length:2919 Min. : 0.0 Length:2919 Length:2919   
## Class :character 1st Qu.: 0.0 Class :character Class :character   
## Mode :character Median : 0.0 Mode :character Mode :character   
## Mean : 102.2   
## 3rd Qu.: 164.0   
## Max. :1600.0   
## NA's :23   
## Foundation BsmtQual BsmtCond   
## Length:2919 Length:2919 Length:2919   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2   
## Length:2919 Length:2919 Min. : 0.0 Length:2919   
## Class :character Class :character 1st Qu.: 0.0 Class :character   
## Mode :character Mode :character Median : 368.5 Mode :character   
## Mean : 441.4   
## 3rd Qu.: 733.0   
## Max. :5644.0   
## NA's :1   
## BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating   
## Min. : 0.00 Min. : 0.0 Min. : 0.0 Length:2919   
## 1st Qu.: 0.00 1st Qu.: 220.0 1st Qu.: 793.0 Class :character   
## Median : 0.00 Median : 467.0 Median : 989.5 Mode :character   
## Mean : 49.58 Mean : 560.8 Mean :1051.8   
## 3rd Qu.: 0.00 3rd Qu.: 805.5 3rd Qu.:1302.0   
## Max. :1526.00 Max. :2336.0 Max. :6110.0   
## NA's :1 NA's :1 NA's :1   
## HeatingQC CentralAir Electrical X1stFlrSF   
## Length:2919 Length:2919 Length:2919 Min. : 334   
## Class :character Class :character Class :character 1st Qu.: 876   
## Mode :character Mode :character Mode :character Median :1082   
## Mean :1160   
## 3rd Qu.:1388   
## Max. :5095   
##   
## X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath   
## Min. : 0.0 Min. : 0.000 Min. : 334 Min. :0.0000   
## 1st Qu.: 0.0 1st Qu.: 0.000 1st Qu.:1126 1st Qu.:0.0000   
## Median : 0.0 Median : 0.000 Median :1444 Median :0.0000   
## Mean : 336.5 Mean : 4.694 Mean :1501 Mean :0.4299   
## 3rd Qu.: 704.0 3rd Qu.: 0.000 3rd Qu.:1744 3rd Qu.:1.0000   
## Max. :2065.0 Max. :1064.000 Max. :5642 Max. :3.0000   
## NA's :2   
## BsmtHalfBath FullBath HalfBath BedroomAbvGr   
## Min. :0.00000 Min. :0.000 Min. :0.0000 Min. :0.00   
## 1st Qu.:0.00000 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:2.00   
## Median :0.00000 Median :2.000 Median :0.0000 Median :3.00   
## Mean :0.06136 Mean :1.568 Mean :0.3803 Mean :2.86   
## 3rd Qu.:0.00000 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:3.00   
## Max. :2.00000 Max. :4.000 Max. :2.0000 Max. :8.00   
## NA's :2   
## KitchenAbvGr KitchenQual TotRmsAbvGrd Functional   
## Min. :0.000 Length:2919 Min. : 2.000 Length:2919   
## 1st Qu.:1.000 Class :character 1st Qu.: 5.000 Class :character   
## Median :1.000 Mode :character Median : 6.000 Mode :character   
## Mean :1.045 Mean : 6.452   
## 3rd Qu.:1.000 3rd Qu.: 7.000   
## Max. :3.000 Max. :15.000   
##   
## Fireplaces FireplaceQu GarageType GarageYrBlt   
## Min. :0.0000 Length:2919 Length:2919 Min. :1895   
## 1st Qu.:0.0000 Class :character Class :character 1st Qu.:1960   
## Median :1.0000 Mode :character Mode :character Median :1979   
## Mean :0.5971 Mean :1978   
## 3rd Qu.:1.0000 3rd Qu.:2002   
## Max. :4.0000 Max. :2207   
## NA's :159   
## GarageFinish GarageCars GarageArea GarageQual   
## Length:2919 Min. :0.000 Min. : 0.0 Length:2919   
## Class :character 1st Qu.:1.000 1st Qu.: 320.0 Class :character   
## Mode :character Median :2.000 Median : 480.0 Mode :character   
## Mean :1.767 Mean : 472.9   
## 3rd Qu.:2.000 3rd Qu.: 576.0   
## Max. :5.000 Max. :1488.0   
## NA's :1 NA's :1   
## GarageCond PavedDrive WoodDeckSF OpenPorchSF   
## Length:2919 Length:2919 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 : 26.00   
## Mean : 93.71 Mean : 47.49   
## 3rd Qu.: 168.00 3rd Qu.: 70.00   
## Max. :1424.00 Max. :742.00   
##   
## EnclosedPorch X3SsnPorch ScreenPorch PoolArea   
## Min. : 0.0 Min. : 0.000 Min. : 0.00 Min. : 0.000   
## 1st Qu.: 0.0 1st Qu.: 0.000 1st Qu.: 0.00 1st Qu.: 0.000   
## Median : 0.0 Median : 0.000 Median : 0.00 Median : 0.000   
## Mean : 23.1 Mean : 2.602 Mean : 16.06 Mean : 2.252   
## 3rd Qu.: 0.0 3rd Qu.: 0.000 3rd Qu.: 0.00 3rd Qu.: 0.000   
## Max. :1012.0 Max. :508.000 Max. :576.00 Max. :800.000   
##   
## PoolQC Fence MiscFeature   
## Length:2919 Length:2919 Length:2919   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## MiscVal MoSold YrSold SaleType   
## Min. : 0.00 Min. : 1.000 Min. :2006 Length:2919   
## 1st Qu.: 0.00 1st Qu.: 4.000 1st Qu.:2007 Class :character   
## Median : 0.00 Median : 6.000 Median :2008 Mode :character   
## Mean : 50.83 Mean : 6.213 Mean :2008   
## 3rd Qu.: 0.00 3rd Qu.: 8.000 3rd Qu.:2009   
## Max. :17000.00 Max. :12.000 Max. :2010   
##   
## SaleCondition SalePrice   
## Length:2919 Min. : 34900   
## Class :character 1st Qu.:129975   
## Mode :character Median :163000   
## Mean :180921   
## 3rd Qu.:214000   
## Max. :755000   
## NA's :1459

# Data Visualization

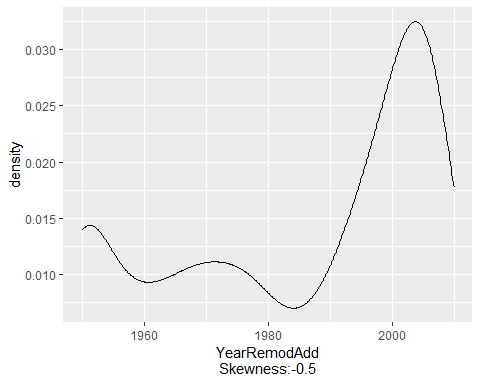
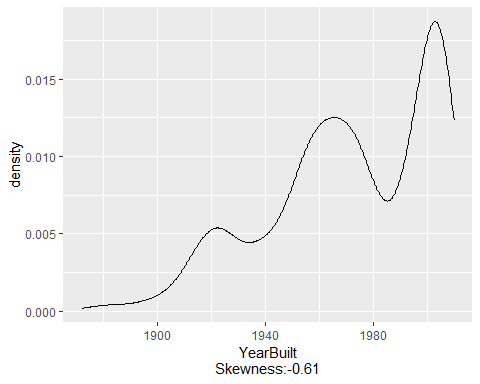
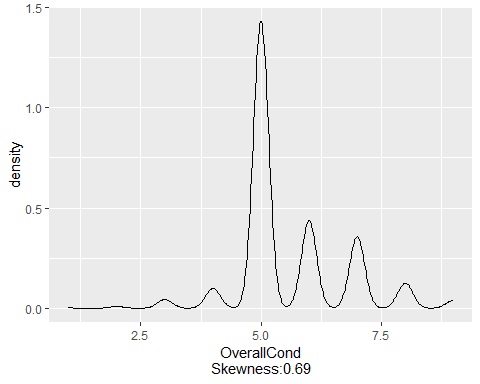
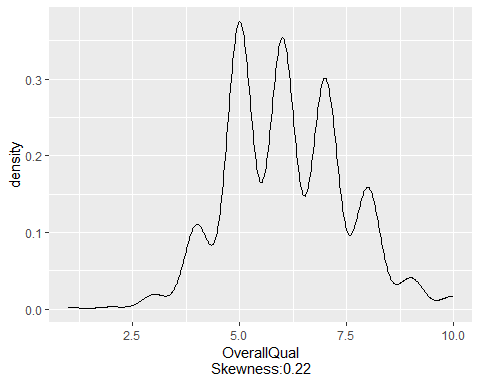
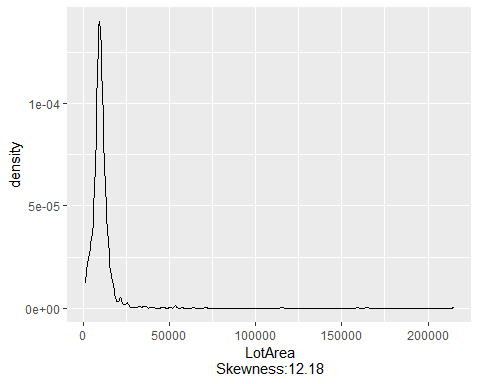
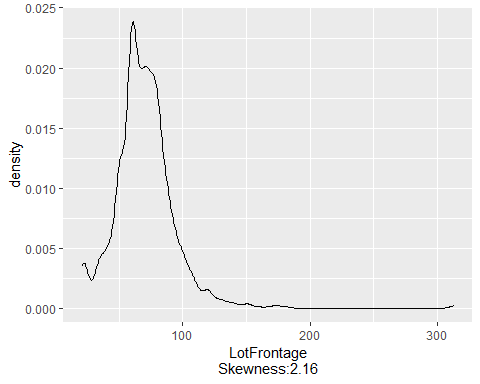
# Creating training dataset with one categorical variable and one numerical variable  
house\_cat\_var <- names(train\_data)[which(sapply(train\_data, is.character))]  
house\_num\_var <- names(train\_data)[which(sapply(train\_data, is.numeric))][-1] #removing ID variable  
  
house\_train\_cat <- train\_data[house\_cat\_var]  
house\_train\_num <- train\_data[house\_num\_var]  
  
#Bar Plot function for count of categorical variables  
par(mfrow=c(22,2))  
plotHist <- function(data) {  
 for (i in colnames(data))  
 {  
 house\_df <- data.frame(x=data[,i])  
 plot <- ggplot(house\_df,aes(x=factor(x),fill=x))+  
 stat\_count()+  
 xlab(colnames(data[i]))+  
 theme(axis.text.x = element\_text(angle = 90))  
 print(plot)  
 }  
}  
plotHist(house\_train\_cat)



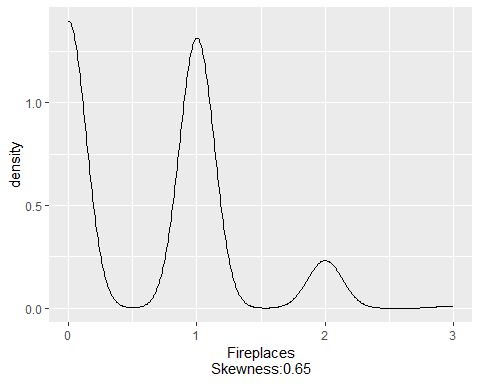
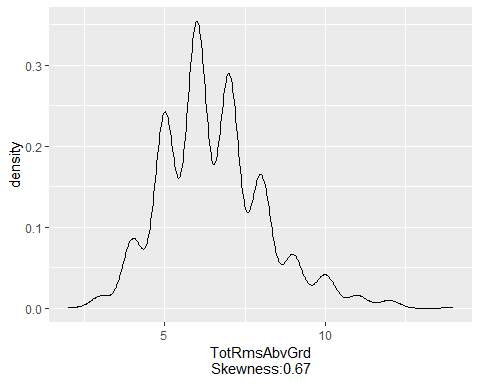
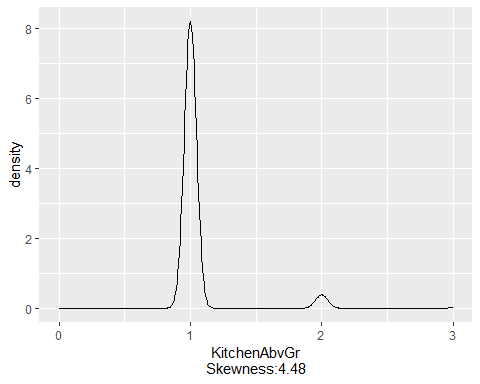
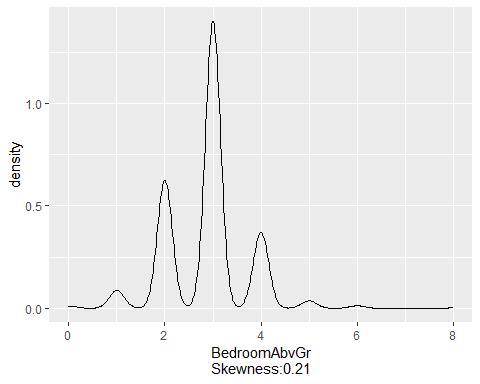
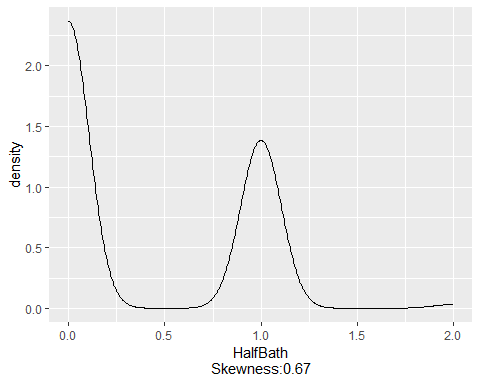
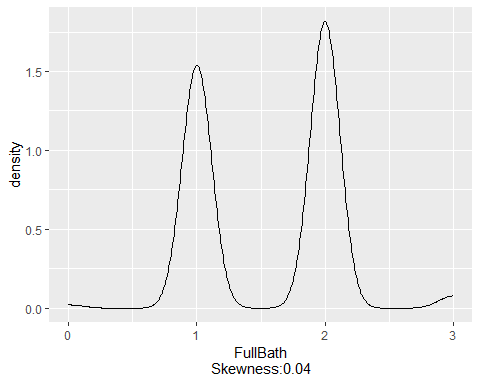
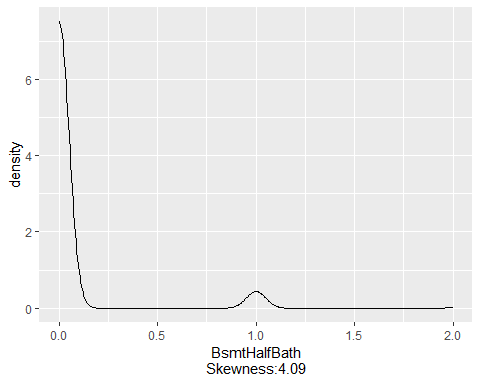
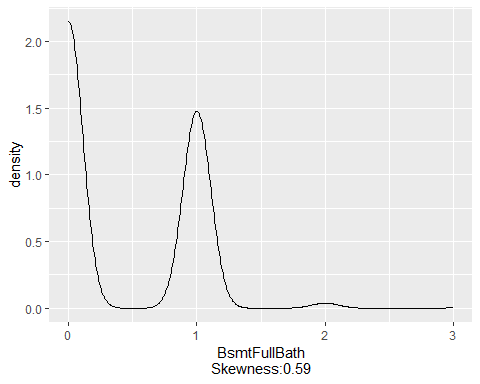
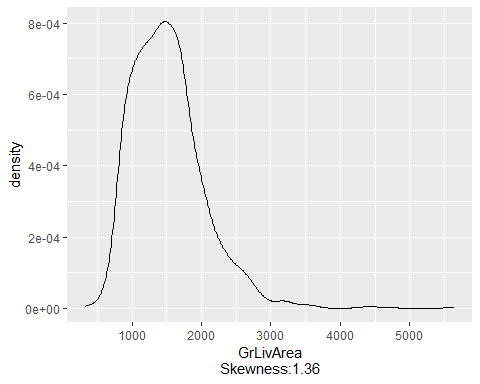
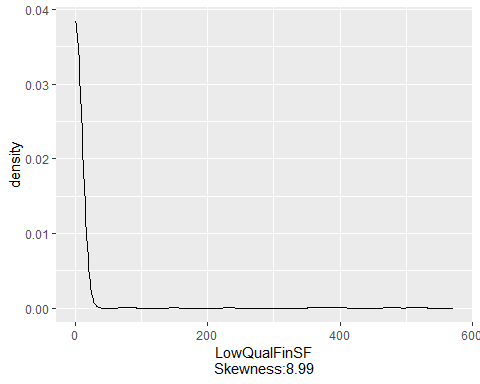
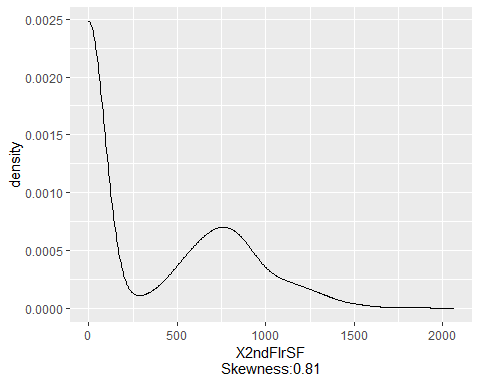
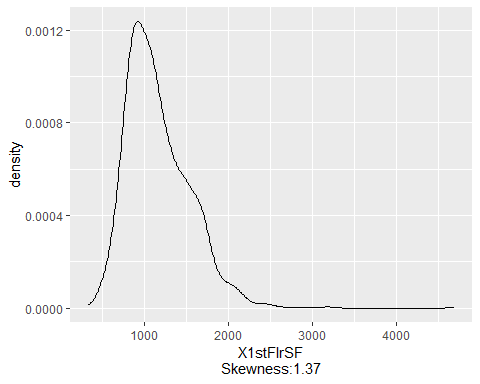
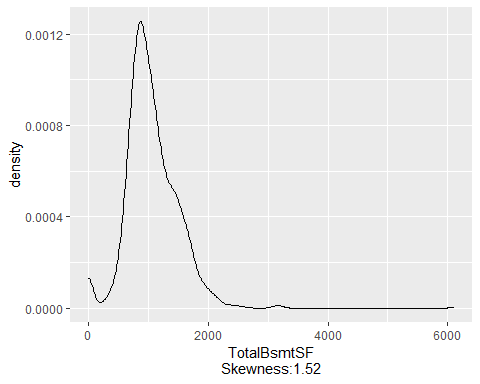
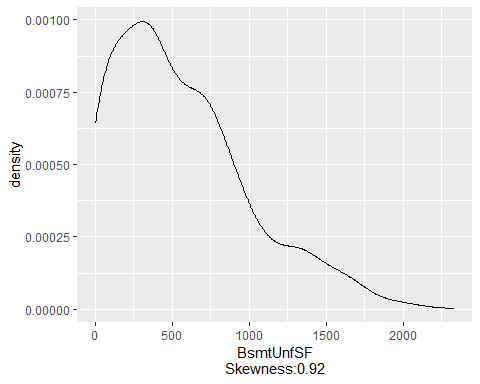
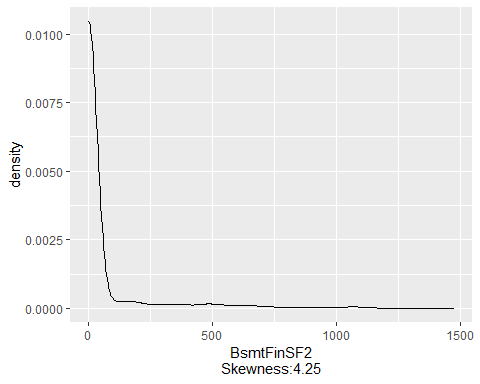
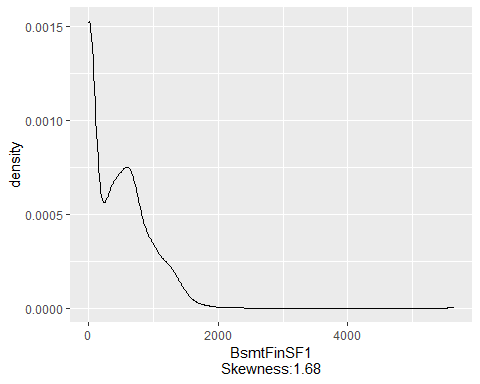
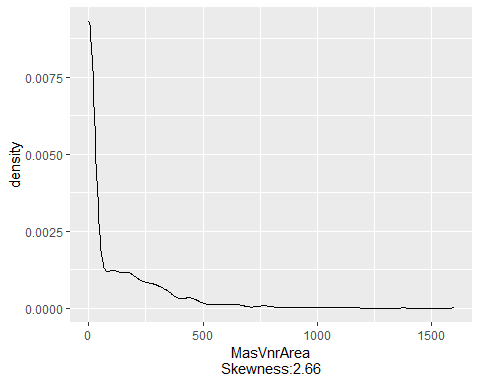
# The histograms above shows that the majority of the houses have 2 full baths, 0 half baths, and have an average of 3 bedrooms.  
  
# Density Plot function for skewness of numeric variables  
plotDensity <- function(data){  
 for (i in colnames(data))  
 {  
 house\_df <- data.frame(x=data[,i])  
 plot <- ggplot(house\_df)+  
 geom\_line(aes(x=x), stat="density")+  
 xlab(paste0(colnames(data[i]), '\n', 'Skewness:', round(skewness(data[,i], na.rm = T), 2)))  
 print(plot)  
 }  
}  
# Calling plot density function for numeric variables  
plotDensity(house\_train\_num)



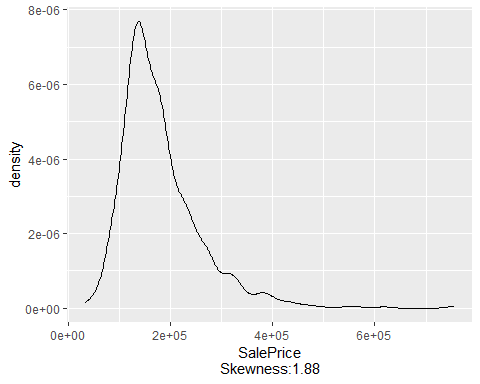
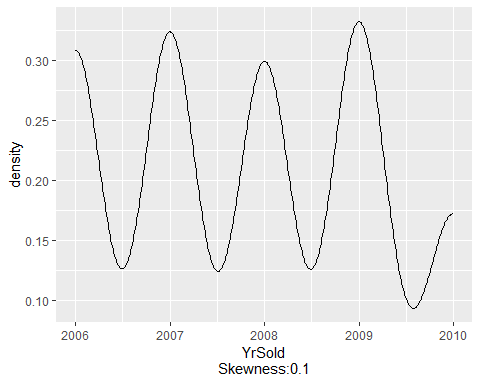
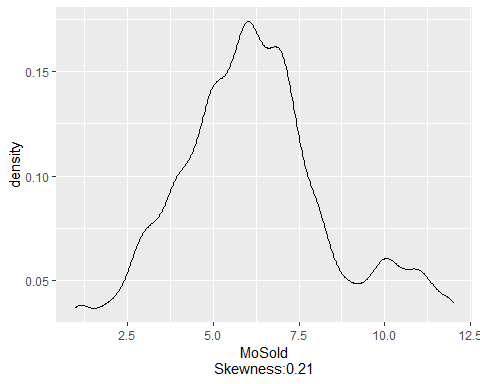
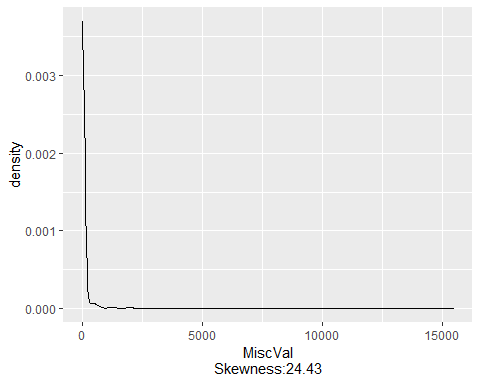
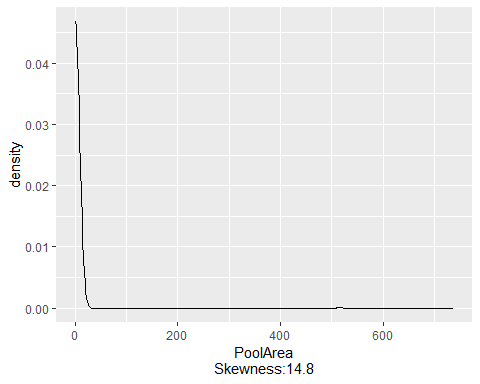
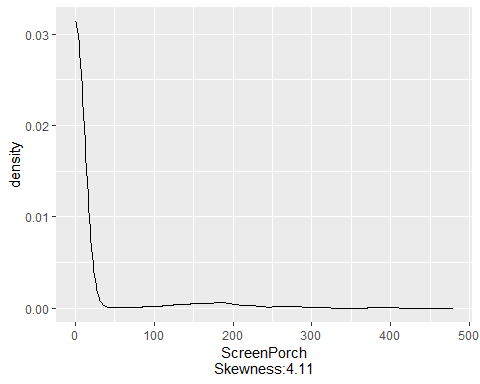
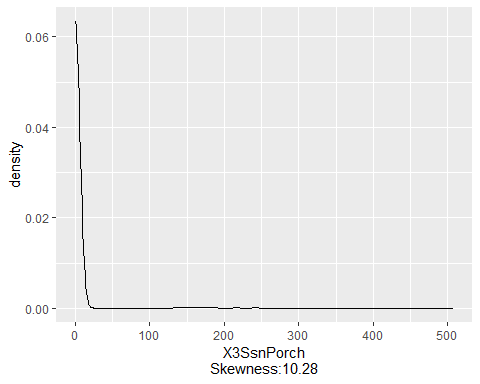
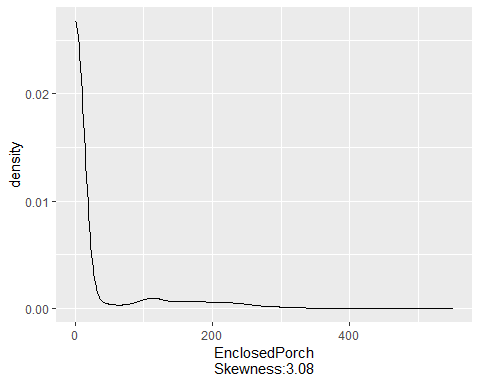
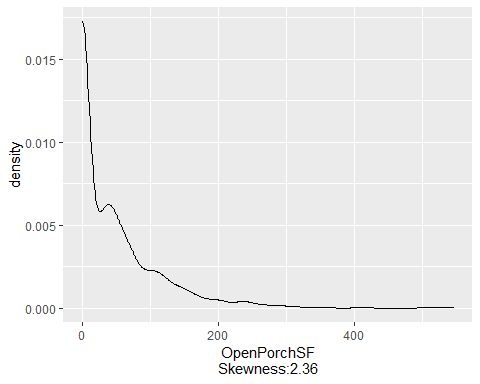
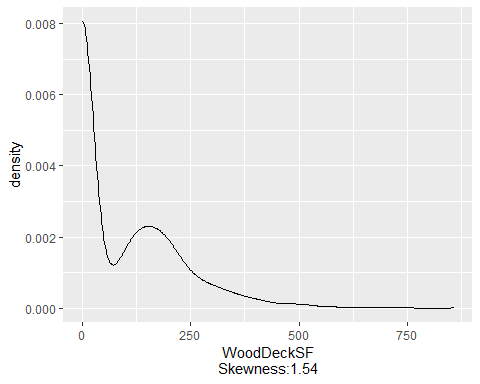
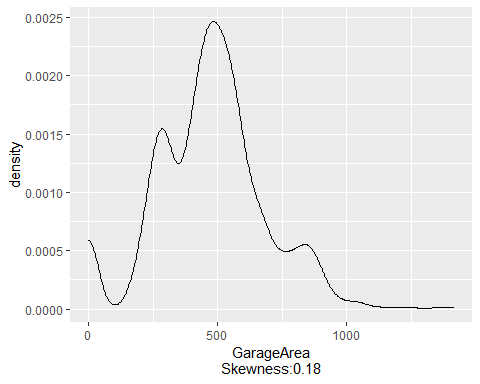
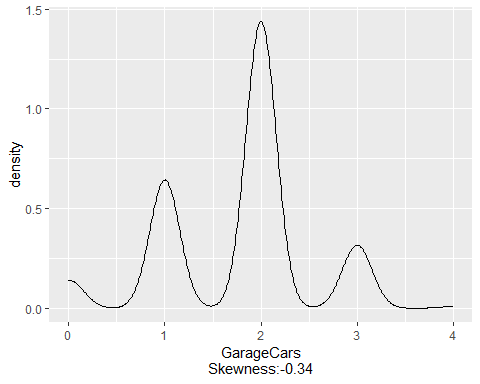
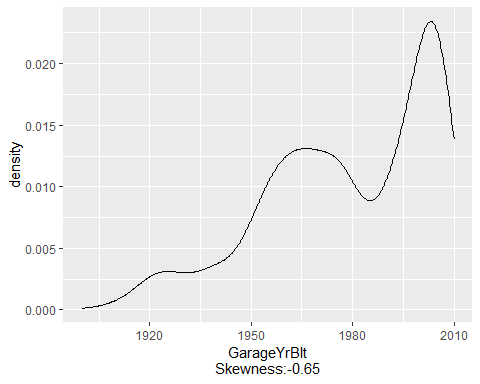
## Warning: Removed 259 rows containing non-finite values (stat\_density).



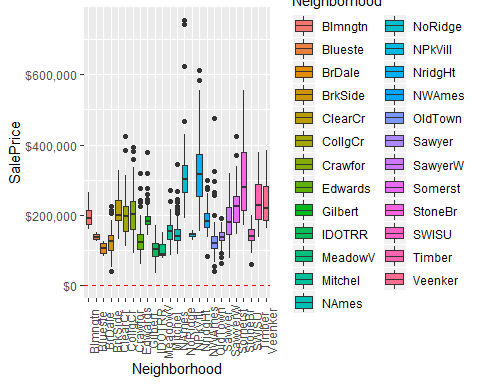
## Warning: Removed 8 rows containing non-finite values (stat\_density).



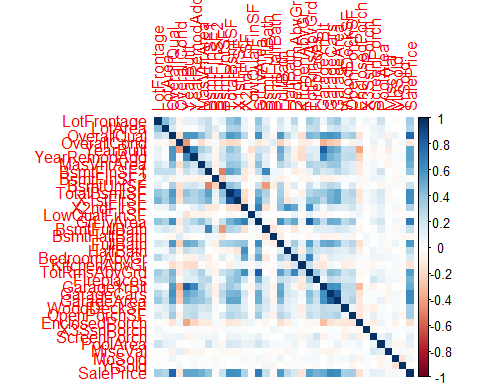
## Warning: Removed 81 rows containing non-finite values (stat\_density).



# Density plots of the features indicates that the features are skewed. The denisty plot for YearBuilt shows that the data set contains a mix of new and old houses. It shows a downturn in the number of houses in recent years, possibily due to the housing crisis  
  
# Box plot for Neighborhood Vs Salesprice  
ggplot(train\_data, aes(x=Neighborhood, y=SalePrice, fill=Neighborhood))+  
 geom\_boxplot()+  
 geom\_hline(aes(yintercept=80), color="red", linetype="dashed")+  
 theme(axis.text.x = element\_text(angle = 90))+  
 scale\_y\_continuous(labels=dollar\_format())



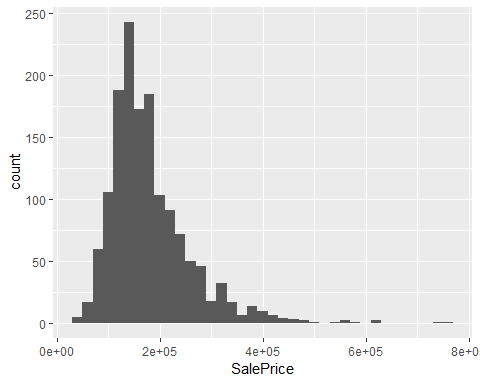
# Boxplot between the neighboorhoods and sale price shows that BrookSide and South & West of Iowa State University have cheap houses. While Northridge and Northridge Heights are rich neighborhoods with several outliers in terms of price.  
  
par(mfrow=c(1,1))  
# Exploring correlation  
cor\_var <- cor(na.omit(house\_train\_num[,-1]))  
corrplot(cor\_var,method = "color")

 #Handling missing values

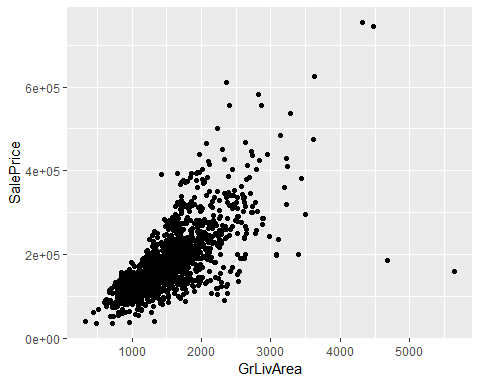
# Distribution of target variable (SalePrice)  
quantile(train\_data$SalePrice)

## 0% 25% 50% 75% 100%   
## 34900 129975 163000 214000 755000

#Histogram for target variable  
ggplot(train\_data)+geom\_histogram(aes(x=SalePrice),binwidth = 20000)



#It is not normal distribution and is positive skew  
  
#Plotting GrLivArea to see if there are any outliers  
ggplot(train\_data,aes(x=GrLivArea,y=SalePrice))+geom\_point()



#Removing outliers from GrLivArea field(>4000)  
train\_data <- train\_data[train\_data$GrLivArea<=4000,]  
  
#To find missing values for all variables in combined dataset(train+test)  
  
#Creating a variable for all missing data  
missing\_ind <- sapply(train\_data, function(x) sum(is.na(x)))  
missing\_data <- data.frame(index=names(train\_data),MissingValues=missing\_ind)  
missing\_data[missing\_data$MissingValues>0,]

## index MissingValues  
## LotFrontage LotFrontage 259  
## Alley Alley 1365  
## MasVnrType MasVnrType 8  
## MasVnrArea MasVnrArea 8  
## BsmtQual BsmtQual 37  
## BsmtCond BsmtCond 37  
## BsmtExposure BsmtExposure 38  
## BsmtFinType1 BsmtFinType1 37  
## BsmtFinType2 BsmtFinType2 38  
## Electrical Electrical 1  
## FireplaceQu FireplaceQu 690  
## GarageType GarageType 81  
## GarageYrBlt GarageYrBlt 81  
## GarageFinish GarageFinish 81  
## GarageQual GarageQual 81  
## GarageCond GarageCond 81  
## PoolQC PoolQC 1451  
## Fence Fence 1176  
## MiscFeature MiscFeature 1402

#Data Preparation  
  
#Imputing missing 'MasVnrArea' by its mean  
house\_data$MasVnrArea[which(is.na(house\_data$MasVnrArea))] <- mean(house\_data$MasVnrArea,na.rm = T)  
  
#Imputing missing 'LotFrontage' by its median  
house\_data$LotFrontage[which(is.na(house\_data$LotFrontage))] <- mean(house\_data$LotFrontage,na.rm = T)  
  
#Changing NA in'GarageBlt' to 0  
house\_data$GarageYrBlt[which(is.na(house\_data$GarageYrBlt))] <- 0  
  
#Changing NA in 'Alley' to None   
house\_data$Alley[which(is.na(house\_data$Alley))] <- "None"  
table(house\_data$Alley)

##   
## Grvl None Pave   
## 120 2721 78

#Changing NA in 'MasVnrType' to None  
house\_data$MasVnrType[which(is.na(house\_data$MasVnrType))] <- "None"  
table(house\_data$MasVnrType)

##   
## BrkCmn BrkFace None Stone   
## 25 879 1766 249

#Changing NA in 'FireplaceQu' to None  
house\_data$FireplaceQu[which(is.na(house\_data$FireplaceQu))] <- "None"  
table(house\_data$FireplaceQu)

##   
## Ex Fa Gd None Po TA   
## 43 74 744 1420 46 592

#Changing NA in 'PoolQc' to None   
house\_data$PoolQC[which(is.na(house\_data$PoolQC))] <- "None"  
table(house\_data$PoolQC)

##   
## Ex Fa Gd None   
## 4 2 4 2909

#Changing NA in 'Fence' to None   
house\_data$Fence[which(is.na(house\_data$Fence))] <- "None"  
table(house\_data$Fence)

##   
## GdPrv GdWo MnPrv MnWw None   
## 118 112 329 12 2348

#Changing NA in 'MiscFeature' to None   
house\_data$MiscFeature[which(is.na(house\_data$MiscFeature))] <- "None"  
table(house\_data$MiscFeature)

##   
## Gar2 None Othr Shed TenC   
## 5 2814 4 95 1

#Changing NA in 'GarageType' to None   
house\_data$GarageType[which(is.na(house\_data$GarageType))] <- "None"  
table(house\_data$GarageType)

##   
## 2Types Attchd Basment BuiltIn CarPort Detchd None   
## 23 1723 36 186 15 779 157

#Changing NA in 'GarageFinish' to None   
house\_data$GarageFinish[which(is.na(house\_data$GarageFinish))] <- "None"  
table(house\_data$GarageFinish)

##   
## Fin None RFn Unf   
## 719 159 811 1230

#Changing NA in 'GarageQual' to None   
house\_data$GarageQual[which(is.na(house\_data$GarageQual))] <- "None"  
table(house\_data$GarageQual)

##   
## Ex Fa Gd None Po TA   
## 3 124 24 159 5 2604

#Changing NA in 'GarageCond' to None   
house\_data$GarageCond[which(is.na(house\_data$GarageCond))] <- "None"  
table(house\_data$GarageCond)

##   
## Ex Fa Gd None Po TA   
## 3 74 15 159 14 2654

#Changing NA in 'BsmtQual' to None   
house\_data$BsmtQual[which(is.na(house\_data$BsmtQual))] <- "None"  
table(house\_data$BsmtQual)

##   
## Ex Fa Gd None TA   
## 258 88 1209 81 1283

#Changing NA in 'BsmtCond' to None   
house\_data$BsmtCond[which(is.na(house\_data$BsmtCond))] <- "None"  
table(house\_data$BsmtCond)

##   
## Fa Gd None Po TA   
## 104 122 82 5 2606

#Changing NA in 'BsmtExposure' to None   
house\_data$BsmtExposure[which(is.na(house\_data$BsmtExposure))] <- "None"  
table(house\_data$BsmtExposure)

##   
## Av Gd Mn No None   
## 418 276 239 1904 82

#Changing NA in 'BsmtFinType1' to None   
house\_data$BsmtFinType1[which(is.na(house\_data$BsmtFinType1))] <- "None"  
table(house\_data$BsmtFinType1)

##   
## ALQ BLQ GLQ LwQ None Rec Unf   
## 429 269 849 154 79 288 851

#Changing NA in 'BsmtFinType2' to None   
house\_data$BsmtFinType2[which(is.na(house\_data$BsmtFinType2))] <- "None"  
table(house\_data$BsmtFinType2)

##   
## ALQ BLQ GLQ LwQ None Rec Unf   
## 52 68 34 87 80 105 2493

#Changing NA in 'Electrical' to None   
house\_data$Electrical[which(is.na(house\_data$Electrical))] <- "None"  
table(house\_data$Electrical)

##   
## FuseA FuseF FuseP Mix None SBrkr   
## 188 50 8 1 1 2671

#Factorizing the variables  
house\_data$Alley <- factor(house\_data$Alley)  
house\_data$MasVnrType <- factor(house\_data$MasVnrType)  
house\_data$FireplaceQu <- factor(house\_data$FireplaceQu)  
house\_data$PoolQC <- factor(house\_data$PoolQC)  
house\_data$Fence <- factor(house\_data$Fence)  
house\_data$MiscFeature <- factor(house\_data$MiscFeature)  
house\_data$GarageType <- factor(house\_data$GarageType)  
house\_data$GarageFinish <- factor(house\_data$GarageFinish)  
house\_data$GarageQual <- factor(house\_data$GarageQual)  
house\_data$GarageCond <- factor(house\_data$GarageCond)  
house\_data$BsmtQual <- factor(house\_data$BsmtQual)  
house\_data$BsmtCond <- factor(house\_data$BsmtCond)  
house\_data$BsmtExposure <- factor(house\_data$BsmtExposure)  
house\_data$BsmtFinType2 <- factor(house\_data$BsmtFinType2)  
house\_data$BsmtFinType1 <- factor(house\_data$BsmtFinType1)  
house\_data$Electrical <- factor(house\_data$Electrical)  
house\_data$MSZoning<- factor(house\_data$MSZoning)  
house\_data$Street <- factor(house\_data$Street)  
house\_data$LotShape <-factor(house\_data$LotShape)  
house\_data$LandContour<-factor(house\_data$LandContour)  
house\_data$Utilities<-factor(house\_data$Utilities)  
house\_data$LotConfig<-factor(house\_data$LotConfig)  
house\_data$LandSlope<-factor(house\_data$LandSlope)  
house\_data$Neighborhood<-factor(house\_data$Neighborhood)  
house\_data$Condition1<-factor(house\_data$Condition1)  
house\_data$Condition2<-factor(house\_data$Condition2)  
house\_data$BldgType<-factor(house\_data$BldgType)  
house\_data$HouseStyle<-factor(house\_data$HouseStyle)  
house\_data$RoofStyle<-factor(house\_data$RoofStyle)  
house\_data$RoofMatl<-factor(house\_data$RoofMatl)  
house\_data$Exterior1st<-factor(house\_data$Exterior1st)  
house\_data$Exterior2nd<-factor(house\_data$Exterior2nd)  
house\_data$ExterQual<-factor(house\_data$ExterQual)  
house\_data$ExterCond<-factor(house\_data$ExterCond)  
house\_data$Foundation<-factor(house\_data$Foundation)  
house\_data$Heating<-factor(house\_data$Heating)  
house\_data$HeatingQC<-factor(house\_data$HeatingQC)  
house\_data$CentralAir<-factor(house\_data$CentralAir)  
house\_data$KitchenQual<-factor(house\_data$KitchenQual)  
house\_data$Functional<-factor(house\_data$Functional)  
house\_data$PavedDrive<-factor(house\_data$PavedDrive)  
house\_data$SaleType<-factor(house\_data$SaleType)  
house\_data$SaleCondition<-factor(house\_data$SaleCondition)  
str(house\_data)

## 'data.frame': 2919 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 : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...  
## $ LotFrontage : num 65 80 68 60 84 ...  
## $ LotArea : int 8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...  
## $ Street : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Alley : Factor w/ 3 levels "Grvl","None",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ LotShape : Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 1 1 1 1 4 1 4 4 ...  
## $ LandContour : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4 4 4 ...  
## $ Utilities : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1 1 ...  
## $ LotConfig : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5 1 5 1 ...  
## $ LandSlope : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14 12 21 17 18 4 ...  
## $ Condition1 : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5 1 1 ...  
## $ Condition2 : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 1 ...  
## $ BldgType : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1 2 ...  
## $ HouseStyle : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6 1 2 ...  
## $ 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 : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ RoofMatl : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Exterior1st : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 13 13 13 7 4 9 ...  
## $ Exterior2nd : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 14 14 14 7 16 9 ...  
## $ MasVnrType : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4 4 3 3 ...  
## $ MasVnrArea : num 196 0 162 0 350 0 186 240 0 0 ...  
## $ ExterQual : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 4 3 4 3 4 4 4 ...  
## $ ExterCond : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...  
## $ Foundation : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...  
## $ BsmtQual : Factor w/ 5 levels "Ex","Fa","Gd",..: 3 3 3 5 3 3 1 3 5 5 ...  
## $ BsmtCond : Factor w/ 5 levels "Fa","Gd","None",..: 5 5 5 2 5 5 5 5 5 5 ...  
## $ BsmtExposure : Factor w/ 5 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4 4 ...  
## $ BsmtFinType1 : Factor w/ 7 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1 7 3 ...  
## $ BsmtFinSF1 : int 706 978 486 216 655 732 1369 859 0 851 ...  
## $ BsmtFinType2 : Factor w/ 7 levels "ALQ","BLQ","GLQ",..: 7 7 7 7 7 7 7 2 7 7 ...  
## $ 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 : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ HeatingQC : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 1 3 1 1 1 1 3 1 ...  
## $ CentralAir : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Electrical : Factor w/ 6 levels "FuseA","FuseF",..: 6 6 6 6 6 6 6 6 2 6 ...  
## $ 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 : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4 4 ...  
## $ TotRmsAbvGrd : int 8 6 6 7 9 5 7 7 8 5 ...  
## $ Functional : Factor w/ 7 levels "Maj1","Maj2",..: 7 7 7 7 7 7 7 7 3 7 ...  
## $ Fireplaces : int 0 1 1 1 1 0 1 2 2 2 ...  
## $ FireplaceQu : Factor w/ 6 levels "Ex","Fa","Gd",..: 4 6 6 3 6 4 3 6 6 6 ...  
## $ GarageType : Factor w/ 7 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2 6 2 ...  
## $ GarageYrBlt : num 2003 1976 2001 1998 2000 ...  
## $ GarageFinish : Factor w/ 4 levels "Fin","None","RFn",..: 3 3 3 4 3 4 3 3 4 3 ...  
## $ 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 : Factor w/ 6 levels "Ex","Fa","Gd",..: 6 6 6 6 6 6 6 6 2 3 ...  
## $ GarageCond : Factor w/ 6 levels "Ex","Fa","Gd",..: 6 6 6 6 6 6 6 6 6 6 ...  
## $ PavedDrive : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...  
## $ 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 : Factor w/ 4 levels "Ex","Fa","Gd",..: 4 4 4 4 4 4 4 4 4 4 ...  
## $ Fence : Factor w/ 5 levels "GdPrv","GdWo",..: 5 5 5 5 5 3 5 5 5 5 ...  
## $ MiscFeature : Factor w/ 5 levels "Gar2","None",..: 2 2 2 2 2 4 2 4 2 2 ...  
## $ 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 : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9 9 9 9 ...  
## $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5 5 1 5 ...  
## $ SalePrice : int 208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...

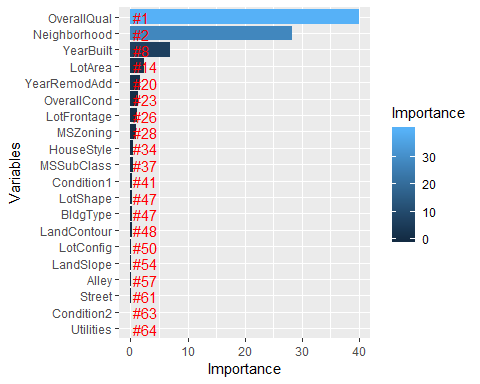
#Taking all the numeric columns in one variable  
col\_class <- sapply(names(house\_data),function(x) {class(house\_data[[x]])})  
numeric\_col <- names(col\_class[col\_class!="factor"])  
  
#Determining skew of each numeric variable  
skew <- sapply(numeric\_col, function(x){skewness(house\_data[[x]],na.rm = T)})  
  
#Transform all variables above threshold  
skew <- skew[skew > 0.75]  
  
#To transform skew variables with log(x+1)  
for(i in names(skew))  
{  
 house\_data[[i]] <- log(house\_data[[i]]+1)  
}

# Model Creation

train\_data <- house\_data[1:1460,]  
test\_data <- house\_data[1461:2919,]  
  
#To make parition in train data  
set.seed(123)  
train\_ind <- sample(1:nrow(train\_data),0.8\*nrow(train\_data))  
new\_train <- train\_data[train\_ind,]  
validate\_train <- train\_data[-train\_ind,]  
new\_train <- subset(new\_train,select=-Id)  
validate\_train <- subset(validate\_train,select=-Id)  
#nrow(new\_train)  
#nrow(validate\_train)  
  
#Building model through random forest  
house\_model <- randomForest(SalePrice~.,data = new\_train)  
print(house\_model)

##   
## Call:  
## randomForest(formula = SalePrice ~ ., data = new\_train)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 26  
##   
## Mean of squared residuals: 0.01827793  
## % Var explained: 88.75

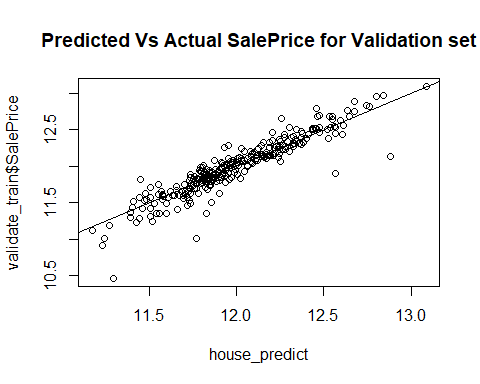
#Get importance  
house\_imp <- importance(house\_model)  
  
#Storing the importance variables in data frame along with IncNodePurity  
varImportance <- data.frame(Variables = row.names(house\_imp),Importance = round(house\_imp[,'IncNodePurity'],2))  
  
#Creating rank variable based on importance  
rankImportance <- varImportance %>% mutate(Rank=paste0("#",dense\_rank(desc(Importance))))  
  
#Visualizing relative importance of first 20 variables  
ggplot(rankImportance[1:20,],aes(x=reorder(Variables,Importance), y=Importance, fill=Importance))+  
 geom\_bar(stat="identity")+  
 geom\_text(aes(x=Variables,y=0.5,label=Rank), hjust=0,vjust=0.55,size=4, color="red")+  
 labs(x="Variables")+  
 coord\_flip()



# predicting using validate set  
house\_predict <- predict(house\_model,validate\_train)  
  
#RMSE function   
RMSE <- function(predicted,actual)  
{  
 rmse <- sqrt(sum((log(predicted)-log(actual))^2)/length(actual))  
 return(rmse)  
}  
  
#Checking RMSE value for validate data  
check\_rmse <- RMSE(house\_predict,validate\_train$SalePrice)  
print(check\_rmse)

## [1] 0.01230236

#Actual Vs Predicted SalePrice for model validation set  
plot(house\_predict,validate\_train$SalePrice,main="Predicted Vs Actual SalePrice for Validation set")  
abline(0,1)



#Antilog function  
antilog <- function(lx,base)   
 {   
 lbx<-lx/log(exp(1),base=base)   
 result<-exp(lbx)   
 result   
}  
  
#Predicting SalePrice for test data and storing into csv file  
house\_test\_predict <- predict(house\_model,test\_data)  
  
#Calling antilog function to store original values in test data set for submission  
output <- data.frame(Id=test\_data$Id,SalePrice=antilog(house\_test\_predict,2.71828))  
write.csv(output,file="submit.csv",row.names = F)