Housing Data Exploratory Analysis

AiO

September 26, 2016

```
library(data.table)
library(testthat)
library(gridExtra)
library(corrplot)
library(GGally)
library(ggplot2)
library(e1071)
library(dplyr)
```

```
## -------------------------------------------------------------------------

## data.table + dplyr code now lives in dtplyr.
## Please library(dtplyr)!

## -------------------------------------------------------------------------

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:GGally':
##
##     nasa

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:testthat':
##
##     matches

## The following objects are masked from 'package:data.table':
##
##     between, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
cat_var <- names(train)[which(sapply(train, is.character))]
cat_car <- c(cat_var, 'BedroomAbvGr', 'HalfBath', ' KitchenAbvGr','BsmtFullBath', 'BsmtHalfBath', 'MSSu
numeric_var <- names(train)[which(sapply(train, is.numeric))]
```

Structure of the data

The housing data set has 1460 rows and 81 features with the target feature Sale Price.

```
dim(train)
```

```
## [1] 1460   81
```

```
str(train)
```

```
## Classes 'data.table' and '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" ...
##  $ 1stFlrSF     : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
##  $ 2ndFlrSF     : 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 ...
##  $ 3SsnPorch    : 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 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Summarize the missing values in the data.

Viewing the first five rows of the data indicates that there are columns which have missing values. The categorical variables with the largest number of missing values are: Alley, FirePlaceQu, PoolQC, Fence, and MiscFeature.

- Alley: indicates the type of alley access
- FirePlaceQu: FirePlace Quality
- PoolQC: Pool Quality
- Fence: Fence Quality
- MiscFeature: Miscellaneous features not covered in other categories

The missing values indicate that majority of the houses do not have alley access, no pool, no fence and no elevator, 2nd garage, shed or tennis court that is covered by the MiscFeature.

The numeric variables do not have as many missing values but there are still some present. There are 259 values for the LotFrontage, 8 missing values for MasVnrArea and 81 missing values for GarageYrBlt.

- LotFrontage: Linear feet of street connected to property
- GarageYrBlt: Year garage was built
- MasVnrArea: Masonry veener area in square feet

Definition of Masonry Veener from google: Veneer masonry is a popular choice for home building and remodeling, because it gives the appearance of a solid brick or stone wall while providing better economy and insulation. It can be used as an addition to conventional wood frame structures, and can be placed on concrete block walls.

Brick veeners are not essential to the stucture of the house but are used to chance the appearance of the wall while providing better insulation. They tend to only have one brick layer.

```r
head(train)
```

```
##    Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape
## 1:  1         60       RL          65    8450   Pave    NA      Reg
## 2:  2         20       RL          80    9600   Pave    NA      Reg
## 3:  3         60       RL          68   11250   Pave    NA      IR1
## 4:  4         70       RL          60    9550   Pave    NA      IR1
## 5:  5         60       RL          84   14260   Pave    NA      IR1
## 6:  6         50       RL          85   14115   Pave    NA      IR1
##    LandContour Utilities LotConfig LandSlope Neighborhood Condition1
## 1:         Lvl    AllPub    Inside       Gtl      CollgCr       Norm
## 2:         Lvl    AllPub       FR2       Gtl      Veenker      Feedr
## 3:         Lvl    AllPub    Inside       Gtl      CollgCr       Norm
## 4:         Lvl    AllPub    Corner       Gtl      Crawfor       Norm
## 5:         Lvl    AllPub       FR2       Gtl      NoRidge       Norm
## 6:         Lvl    AllPub    Inside       Gtl      Mitchel       Norm
##    Condition2 BldgType HouseStyle OverallQual OverallCond YearBuilt
## 1:       Norm     1Fam     2Story           7           5      2003
## 2:       Norm     1Fam     1Story           6           8      1976
## 3:       Norm     1Fam     2Story           7           5      2001
## 4:       Norm     1Fam     2Story           7           5      1915
## 5:       Norm     1Fam     2Story           8           5      2000
## 6:       Norm     1Fam     1.5Fin           5           5      1993
##    YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType
## 1:         2003     Gable  CompShg     VinylSd     VinylSd    BrkFace
## 2:         1976     Gable  CompShg     MetalSd     MetalSd       None
## 3:         2002     Gable  CompShg     VinylSd     VinylSd    BrkFace
## 4:         1970     Gable  CompShg     Wd Sdng     Wd Shng       None
## 5:         2000     Gable  CompShg     VinylSd     VinylSd    BrkFace
## 6:         1995     Gable  CompShg     VinylSd     VinylSd       None
##    MasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond
## 1:        196        Gd        TA      PConc       Gd       TA
## 2:          0        TA        TA     CBlock       Gd       TA
## 3:        162        Gd        TA      PConc       Gd       TA
## 4:          0        TA        TA     BrkTil       TA       Gd
## 5:        350        Gd        TA      PConc       Gd       TA
## 6:          0        TA        TA       Wood       Gd       TA
##    BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2 BsmtUnfSF
## 1:           No          GLQ        706          Unf          0       150
## 2:           Gd          ALQ        978          Unf          0       284
## 3:           Mn          GLQ        486          Unf          0       434
## 4:           No          ALQ        216          Unf          0       540
## 5:           Av          GLQ        655          Unf          0       490
## 6:           No          GLQ        732          Unf          0        64
##    TotalBsmtSF Heating HeatingQC CentralAir Electrical 1stFlrSF 2ndFlrSF
## 1:         856    GasA        Ex          Y      SBrkr      856      854
## 2:        1262    GasA        Ex          Y      SBrkr     1262        0
## 3:         920    GasA        Ex          Y      SBrkr      920      866
## 4:         756    GasA        Gd          Y      SBrkr      961      756
## 5:        1145    GasA        Ex          Y      SBrkr     1145     1053
## 6:         796    GasA        Ex          Y      SBrkr      796      566
##    LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
## 1:            0      1710            1            0        2        1
## 2:            0      1262            0            1        2        0
## 3:            0      1786            1            0        2        1
## 4:            0      1717            1            0        1        0
## 5:            0      2198            1            0        2        1
## 6:            0      1362            1            0        1        1
##    BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
## 1:            3            1          Gd            8        Typ
## 2:            3            1          TA            6        Typ
## 3:            3            1          Gd            6        Typ
## 4:            3            1          Gd            7        Typ
## 5:            4            1          Gd            9        Typ
## 6:            1            1          TA            5        Typ
##    Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
## 1:          0          NA     Attchd        2003          RFn          2
## 2:          1          TA     Attchd        1976          RFn          2
## 3:          1          TA     Attchd        2001          RFn          2
## 4:          1          Gd     Detchd        1998          Unf          3
## 5:          1          TA     Attchd        2000          RFn          3
## 6:          0          NA     Attchd        1993          Unf          2
##    GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
## 1:        548         TA         TA          Y          0          61
## 2:        460         TA         TA          Y        298           0
## 3:        608         TA         TA          Y          0          42
## 4:        642         TA         TA          Y          0          35
## 5:        836         TA         TA          Y        192          84
## 6:        480         TA         TA          Y         40          30
##    EnclosedPorch 3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature
## 1:             0         0           0        0     NA    NA          NA
## 2:             0         0           0        0     NA    NA          NA
## 3:             0         0           0        0     NA    NA          NA
## 4:           272         0           0        0     NA    NA          NA
## 5:             0         0           0        0      NA    NA           NA
## 6:             0       320           0        0      NA MnPrv         Shed
##    MiscVal MoSold YrSold SaleType SaleCondition SalePrice
## 1:       0      2   2008       WD        Normal    208500
## 2:       0      5   2007       WD        Normal    181500
## 3:       0      9   2008       WD        Normal    223500
## 4:       0      2   2006       WD       Abnorml    140000
## 5:       0     12   2008       WD        Normal    250000
## 6:     700     10   2009       WD        Normal    143000
```

```
colSums(sapply(train, is.na))
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0           259             0
##        Street         Alley      LotShape   LandContour     Utilities
##             0          1369             0             0             0
##     LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##             0             0             0             0             0
##      BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##             0             0             0             0             0
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##             0             0             0             0             0
##    MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##             8             8             0             0             0
##      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##            37            37            38            37             0
##  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
##            38             0             0             0             0
##     HeatingQC    CentralAir    Electrical      1stFlrSF      2ndFlrSF
##             0             0             1             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             0             0             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             0             0
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##             0             0           690            81            81
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##            81             0             0            81            81
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch     3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##             0             0          1453          1179          1406
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             0             0
##     SalePrice
##             0
```

```
colSums(sapply(train[,.SD, .SDcols = cat_var], is.na))
```

```
##      MSZoning        Street         Alley      LotShape   LandContour
##             0             0          1369             0             0
##     Utilities     LotConfig     LandSlope  Neighborhood    Condition1
##             0             0             0             0             0
##    Condition2      BldgType    HouseStyle     RoofStyle      RoofMatl
##             0             0             0             0             0
##   Exterior1st   Exterior2nd    MasVnrType     ExterQual     ExterCond
##             0             0             8             0             0
##    Foundation      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1
##             0            37            37            38            37
##  BsmtFinType2       Heating     HeatingQC    CentralAir    Electrical
##            38             0             0             0             1
##   KitchenQual    Functional   FireplaceQu    GarageType  GarageFinish
##             0             0           690            81            81
##    GarageQual    GarageCond    PavedDrive        PoolQC         Fence
##            81            81             0          1453          1179
##   MiscFeature      SaleType SaleCondition
##          1406             0             0
```

```
colSums(sapply(train[,.SD, .SDcols = numeric_var], is.na))
```

```
##            Id    MSSubClass   LotFrontage       LotArea   OverallQual
##             0             0           259             0             0
##   OverallCond     YearBuilt  YearRemodAdd    MasVnrArea    BsmtFinSF1
##             0             0             0             8             0
##    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF      1stFlrSF      2ndFlrSF
##             0             0             0             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             0             0             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd    Fireplaces
##             0             0             0             0             0
##   GarageYrBlt    GarageCars    GarageArea    WoodDeckSF   OpenPorchSF
##            81             0             0             0             0
## EnclosedPorch     3SsnPorch   ScreenPorch      PoolArea       MiscVal
##             0             0             0             0             0
##        MoSold        YrSold     SalePrice
##             0             0             0
```

Let's gain some insight on the number of houses that were remodeled. According to the data dictionary, if the YearBuilt date is different from the YearRemodAdd date then the house was remodeled. Comparing these two rows indicates that 696 houses were remodeled and 764 houses were not remodeled.

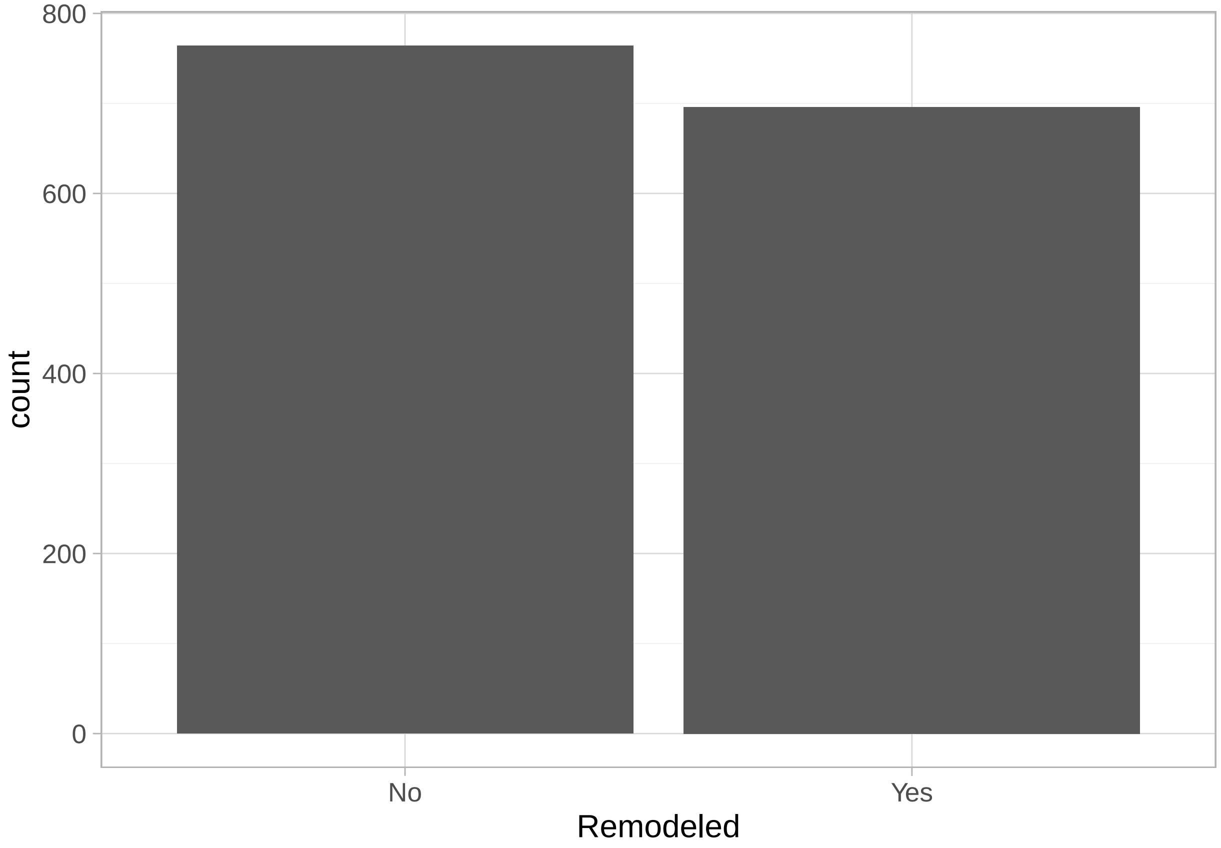
```
sum(train[,'YearRemodAdd', with = FALSE] != train[,'YearBuilt', with = FALSE])
```

```
## [1] 696
```

```
cat('Percentage of houses remodeled',sum(train[,'YearRemodAdd', with = FALSE] != train[,'YearBuilt', wi
```

```
## Percentage of houses remodeled 0.4767123
```

```
train %>% select(YearBuilt, YearRemodAdd) %>%    mutate(Remodeled = as.integer(YearBuilt != YearRemodAd
```



Summarize the numeric values and the structure of the data.

```
summary(train[,.SD, .SDcols =numeric_var])
```

```
##        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        1stFlrSF       2ndFlrSF     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      3SsnPorch     
##  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
##
```

```
cat('Train has', dim(train)[1], 'rows and', dim(train)[2], 'columns.')
```

```
## Train has 1460 rows and 81 columns.
```

```
cat('Test has', dim(test)[1], 'rows and', dim(test)[2], ' columns.')
```

```
## Test has 1459 rows and 80  columns.
```

```
# The percentage of data missing in train.
sum(is.na(train)) / (nrow(train) *ncol(train))
```

```
## [1] 0.05889565
```

```
# The percentage of data missing in test.
sum(is.na(test)) / (nrow(test) * ncol(test))
```

```
## [1] 0.05997258
```

```
# Check for duplicated rows.

cat("The number of duplicated rows are", nrow(train) - nrow(unique(train)))
```

```
## The number of duplicated rows are 0
```

```
####Convert character to factors

train[,(cat_var) := lapply(.SD, as.factor), .SDcols = cat_var]
```

```
train_cat <- train[,.SD, .SDcols = cat_var]
train_cont <- train[,.SD,.SDcols = numeric_var]

plotHist <- function(data_in, i) {
  data <- data.frame(x=data_in[[i]])
  p <- ggplot(data=data, aes(x=factor(x))) + stat_count() + xlab(colnames(data_in)[i]) + theme_light() 
    theme(axis.text.x = element_text(angle = 90, hjust =1))
  return (p)
}

doPlots <- function(data_in, fun, ii, ncol=3) {
  pp <- list()
  for (i in ii) {
    p <- fun(data_in=data_in, i=i)
    pp <- c(pp, list(p))
  }
  do.call("grid.arrange", c(pp, ncol=ncol))
}


plotDen <- function(data_in, i){
  data <- data.frame(x=data_in[[i]], SalePrice = data_in$SalePrice)
  p <- ggplot(data= data) + geom_line(aes(x = x), stat = 'density', size = 1,alpha = 1.0) +
    xlab(paste0((colnames(data_in)[i]), '\n', 'Skewness: ',round(skewness(data_in[[i]], na.rm = TRUE), 2
  return(p)

}
```

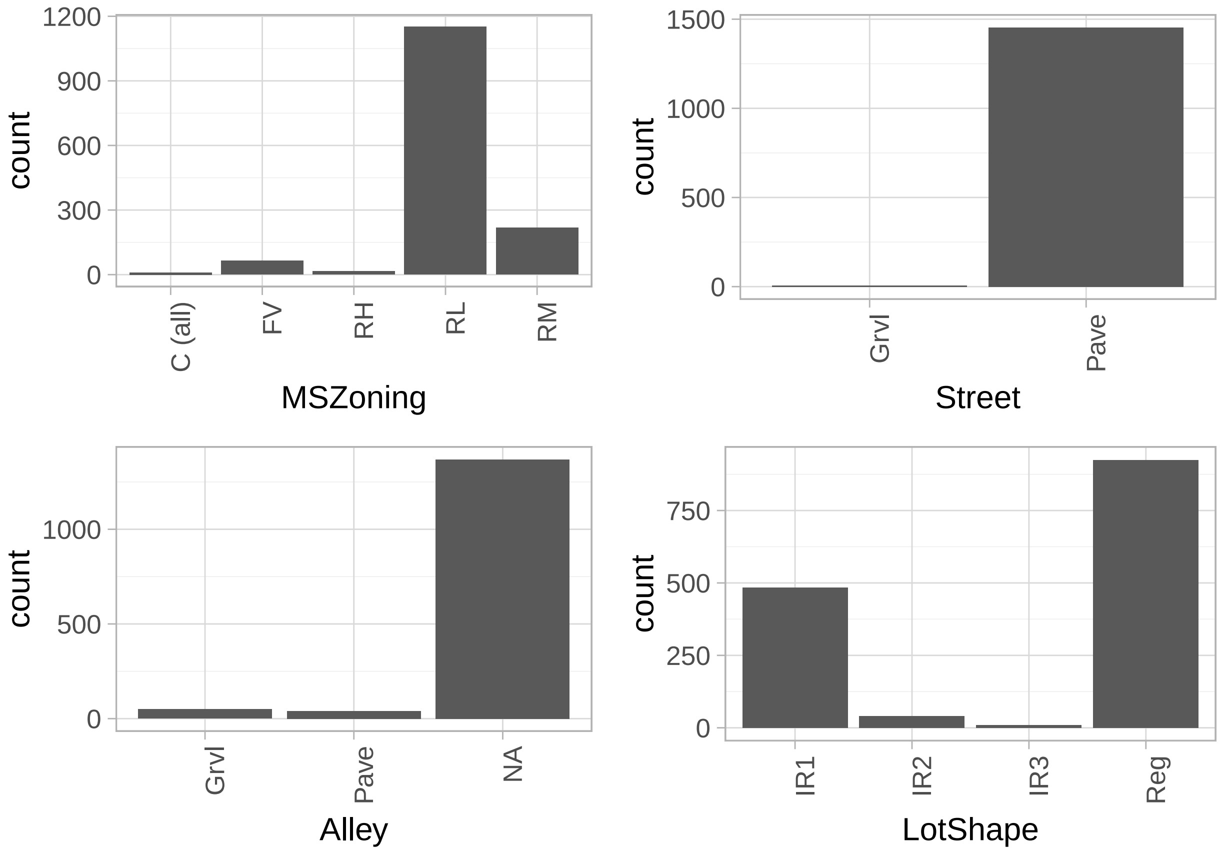
Barplots for the categorical features

The bar plots below offer more insight into the data. MSZoning: bar plot indicates that majority of the houses are located in low density residential areas and medium density residential area.

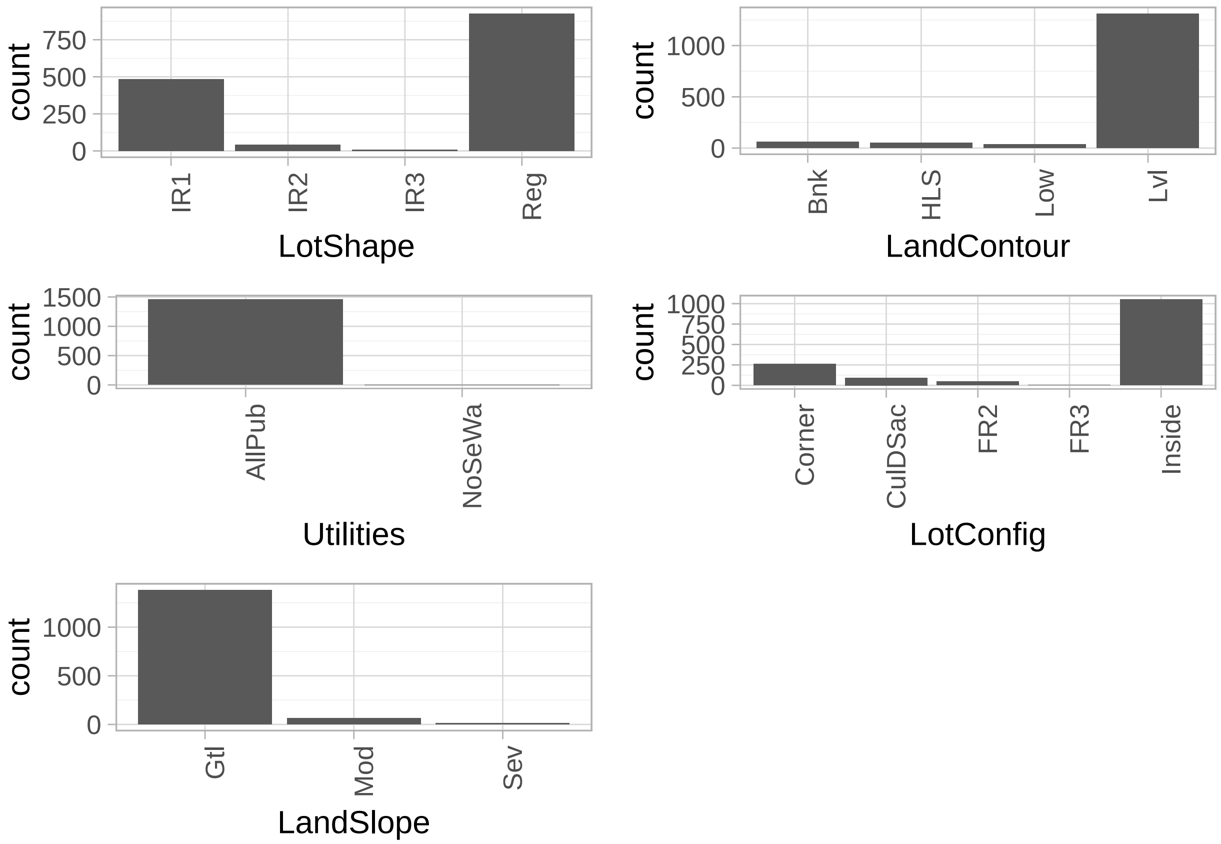
The type of road access to the property tends to be paved and the houses do not have alleys.

- Landcontour: the houses are built on flat properties
- Utilities: Almost all homes have all public utilities (E,G,W, & S)
- LandSlope: most of the properties have a gentle slope

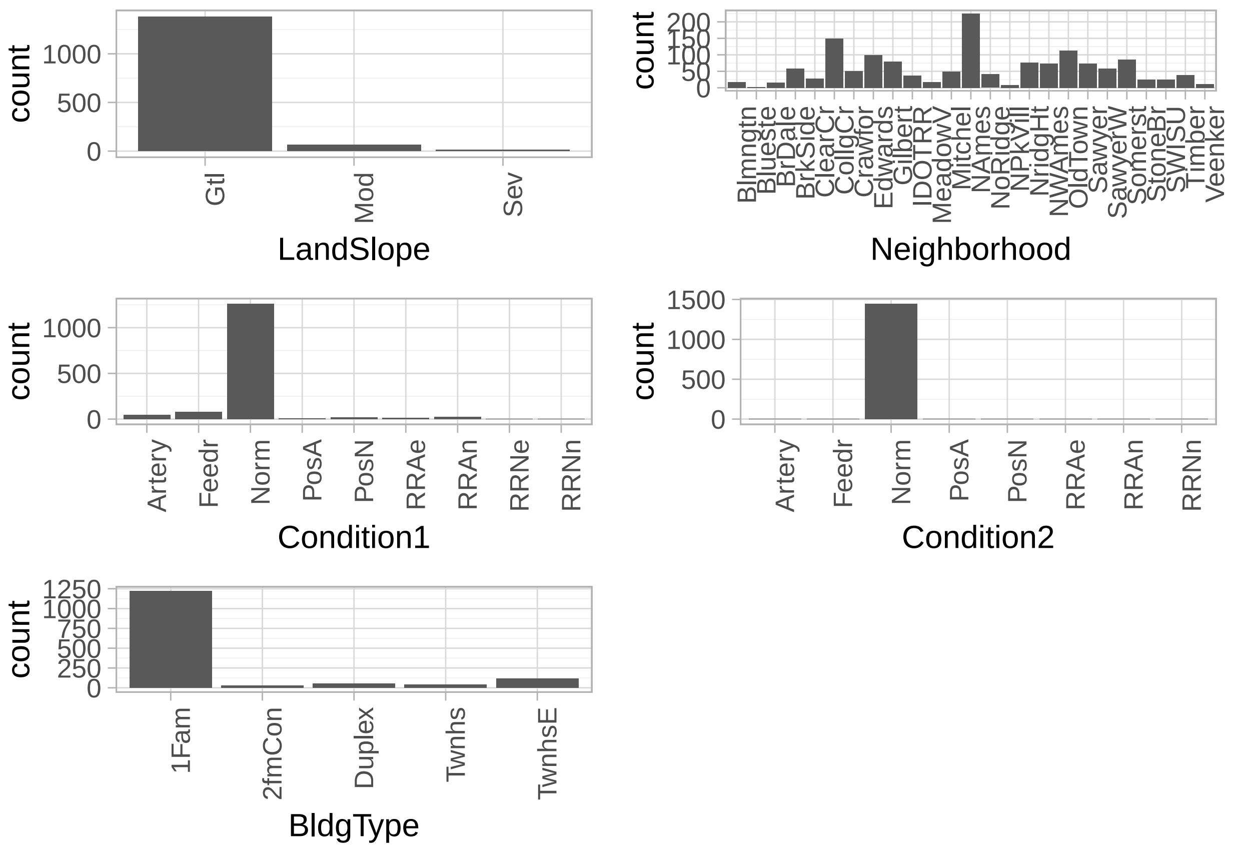
```
doPlots(train_cat, fun = plotHist, ii = 1:4, ncol = 2)
```



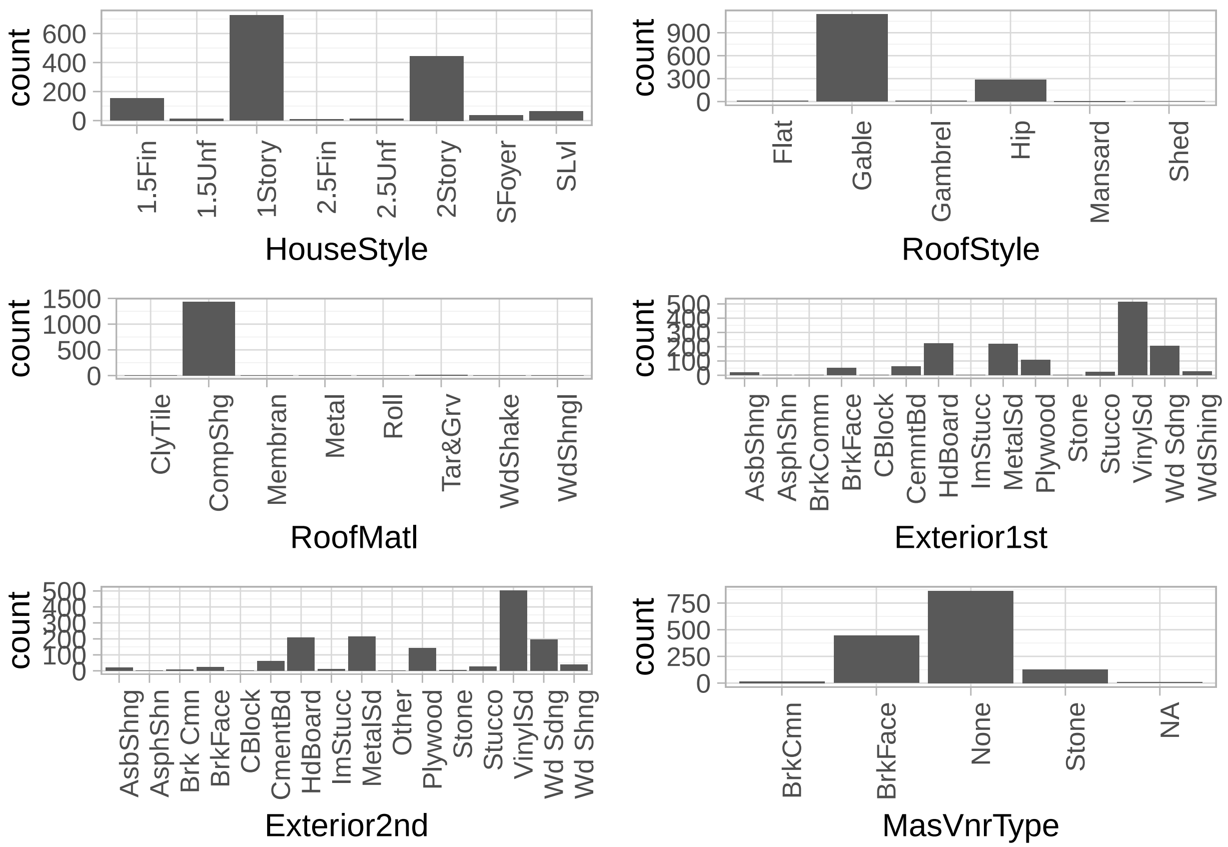
```r
doPlots(train_cat, fun = plotHist, ii  = 4:8, ncol = 2)
```



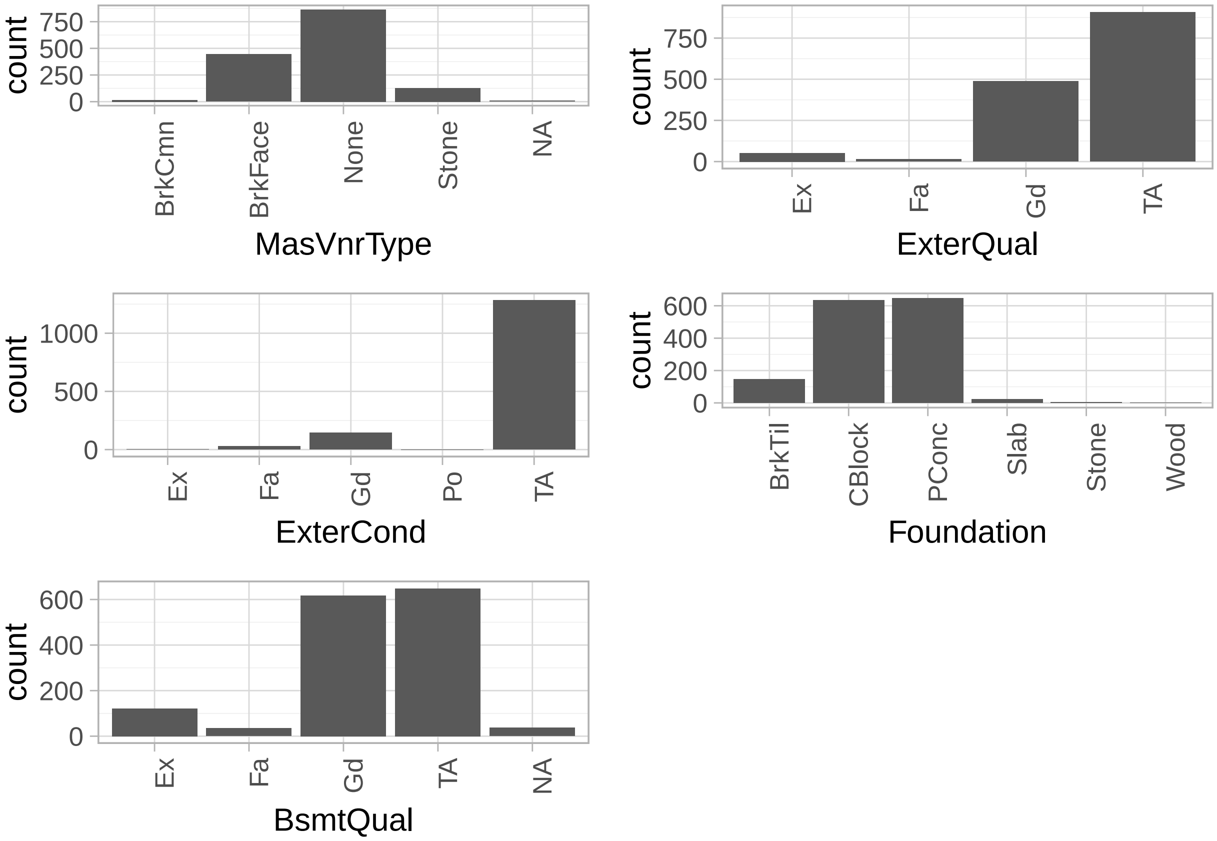
```r
doPlots(train_cat, fun = plotHist, ii = 8:12, ncol = 2)
```



```r
doPlots(train_cat, fun = plotHist, ii = 13:18, ncol = 2)
```

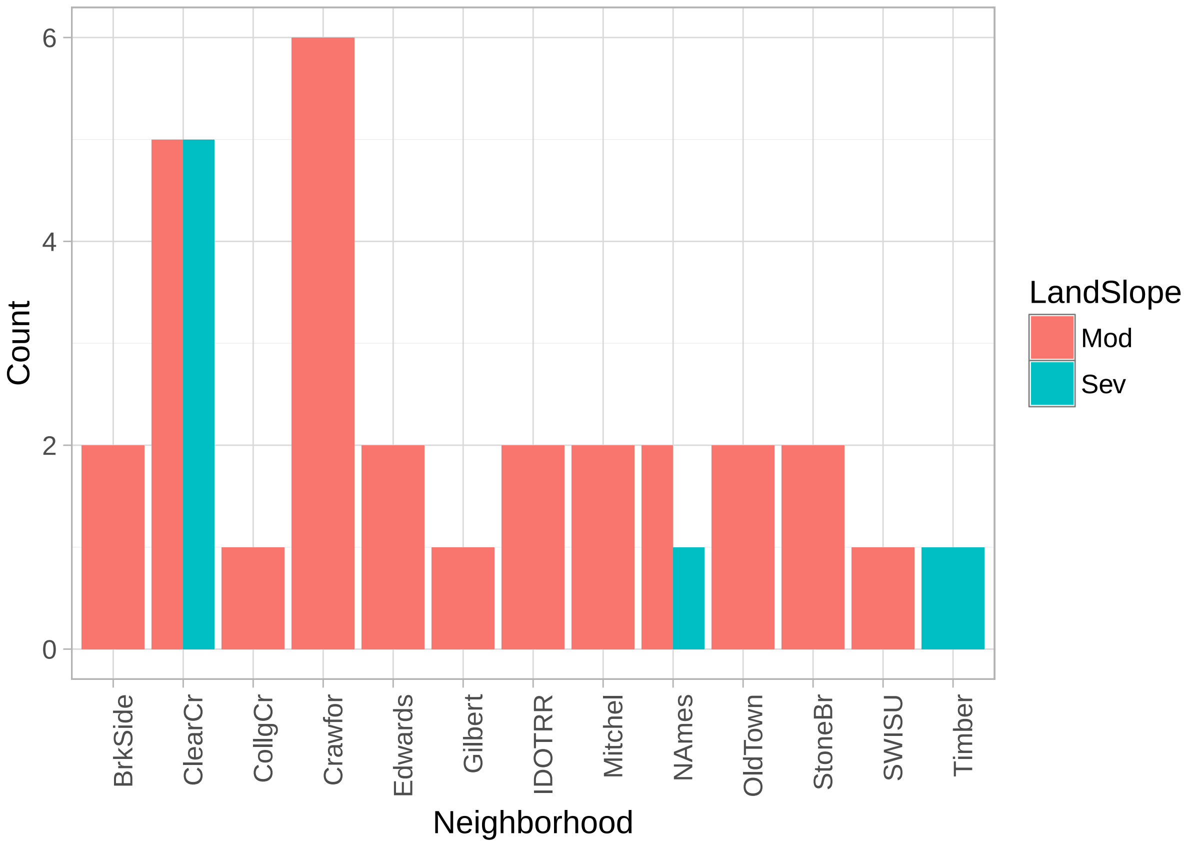


```
doPlots(train_cat, fun = plotHist, ii = 18:22, ncol = 2)
```



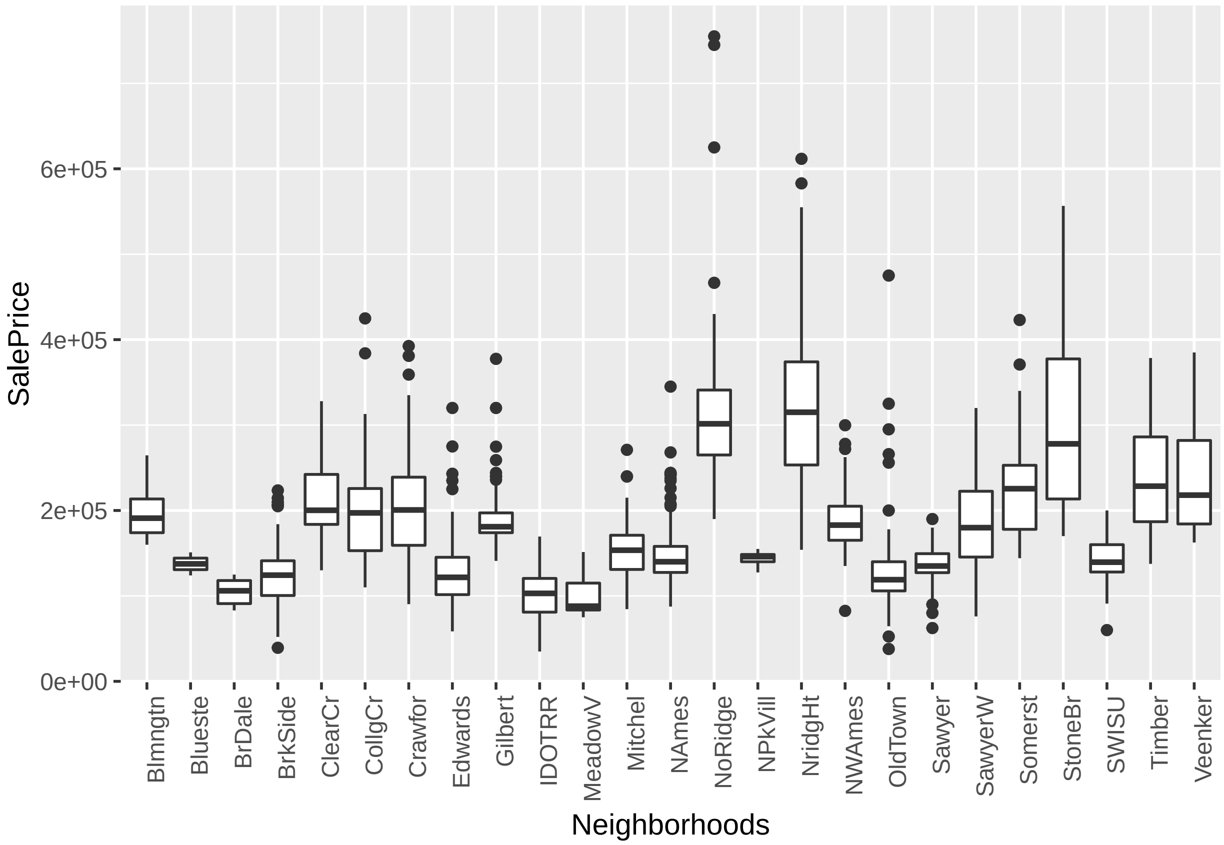
The houses that have sever landslope are located in the Clear Creek and Timberland. The houses with moderate landslope are present in more neighborhood. The Clear Creek and the Crawford neighborhoods seem to have high slopes.

```
train %>% select(LandSlope, Neighborhood, SalePrice) %>% filter(LandSlope == c('Sev', 'Mod')) %>% arrang
```



Plotting a 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.

```
train %>% select(Neighborhood, SalePrice) %>% ggplot(aes(factor(Neighborhood), SalePrice)) + geom_boxpl
```

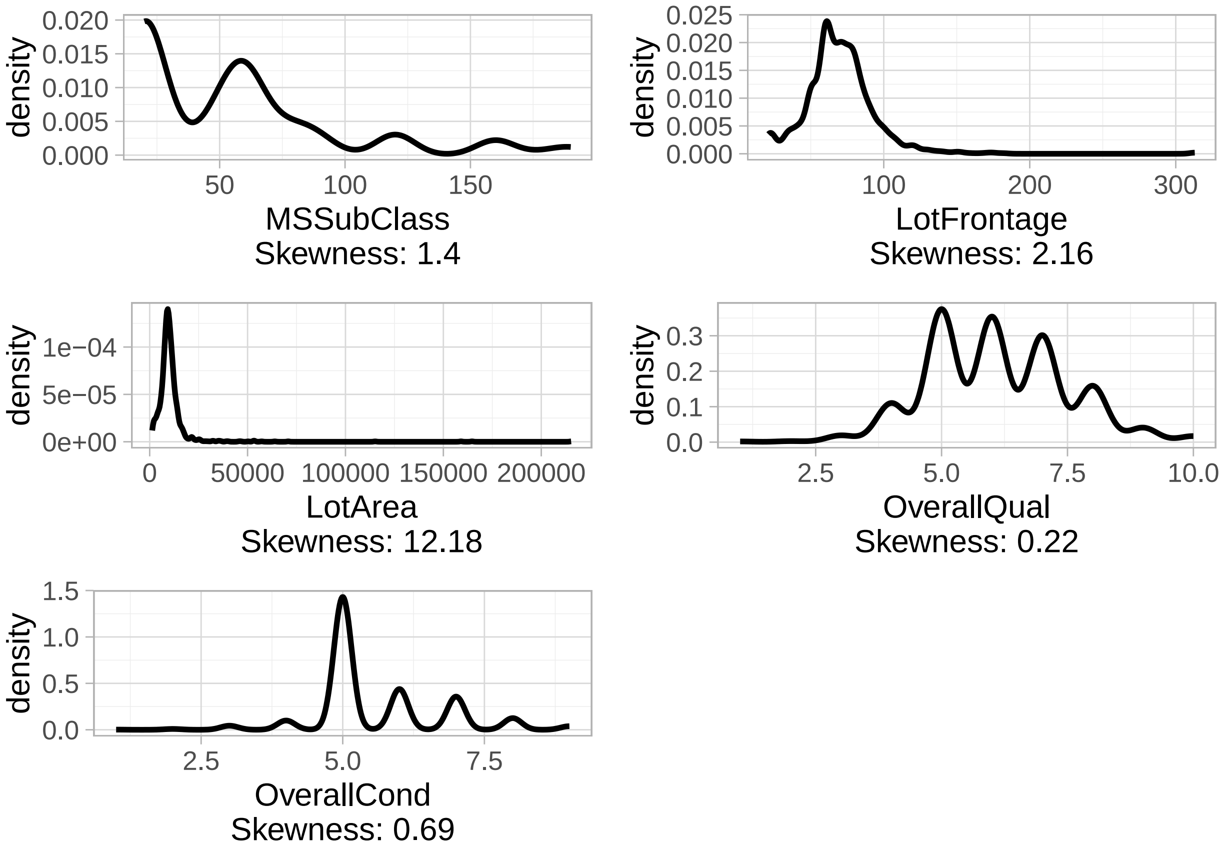


Density plots for numeric variables.

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.

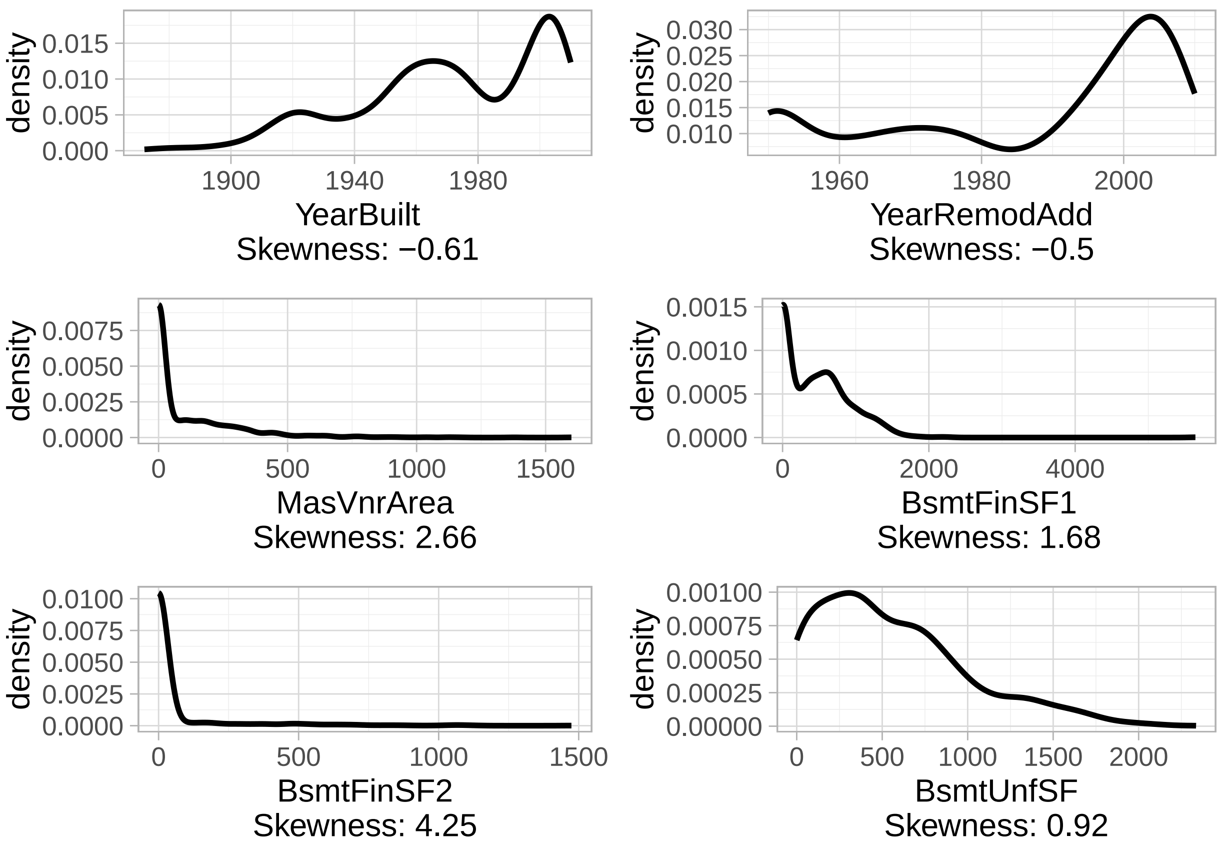
```
doPlots(train_cont, fun = plotDen, ii = 2:6, ncol = 2)
```

```
## Warning: Removed 259 rows containing non-finite values (stat_density).
```

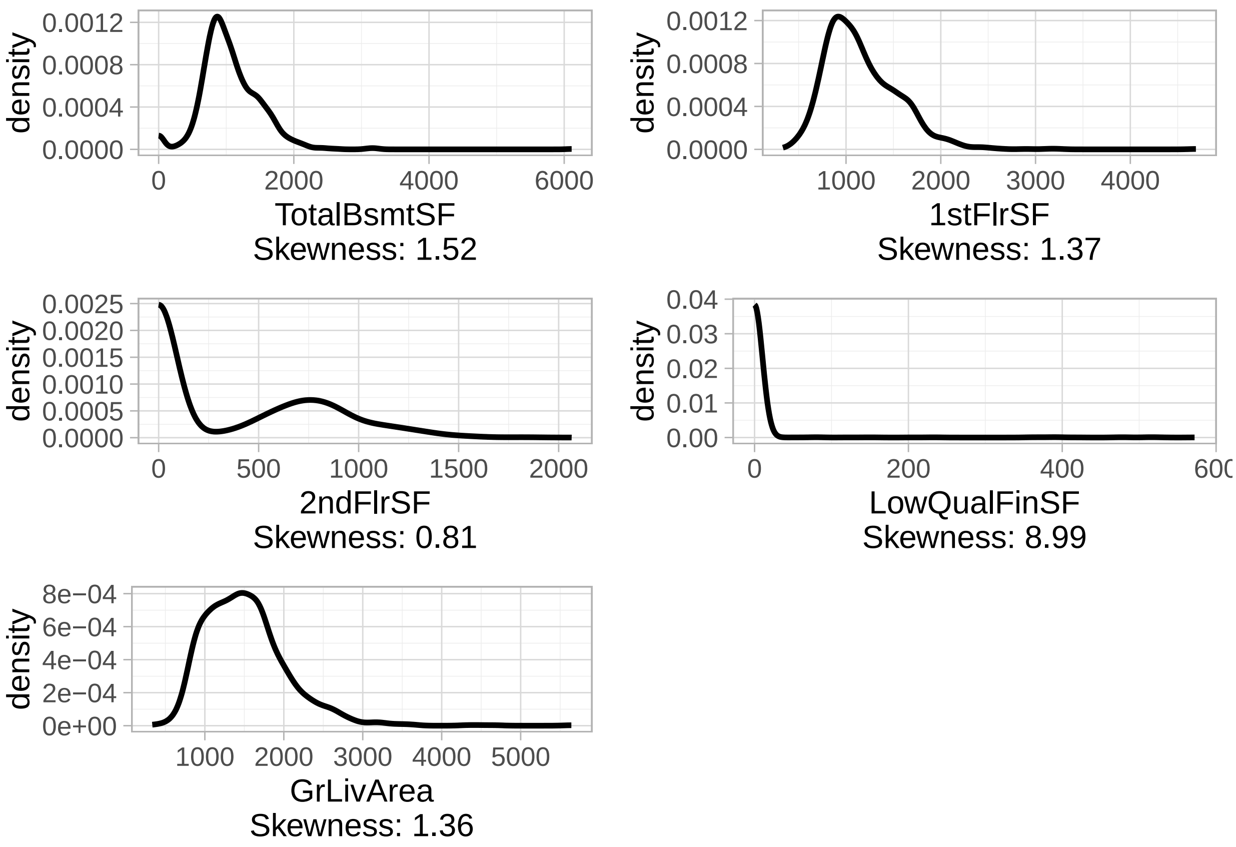


```r
doPlots(train_cont, fun = plotDen, ii = 7:12, ncol = 2)
```

```
## Warning: Removed 8 rows containing non-finite values (stat_density).
```

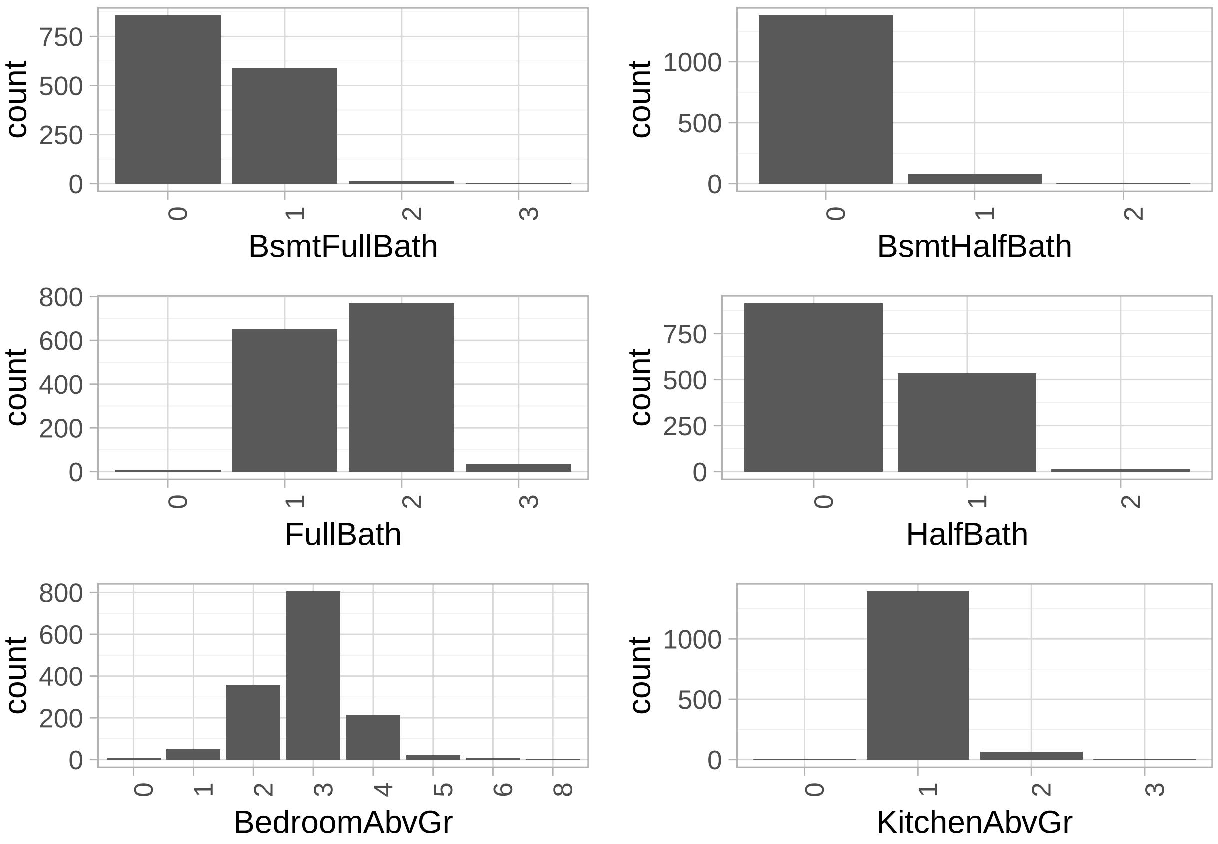


```
doPlots(train_cont, fun = plotDen, ii = 13:17, ncol = 2)
```



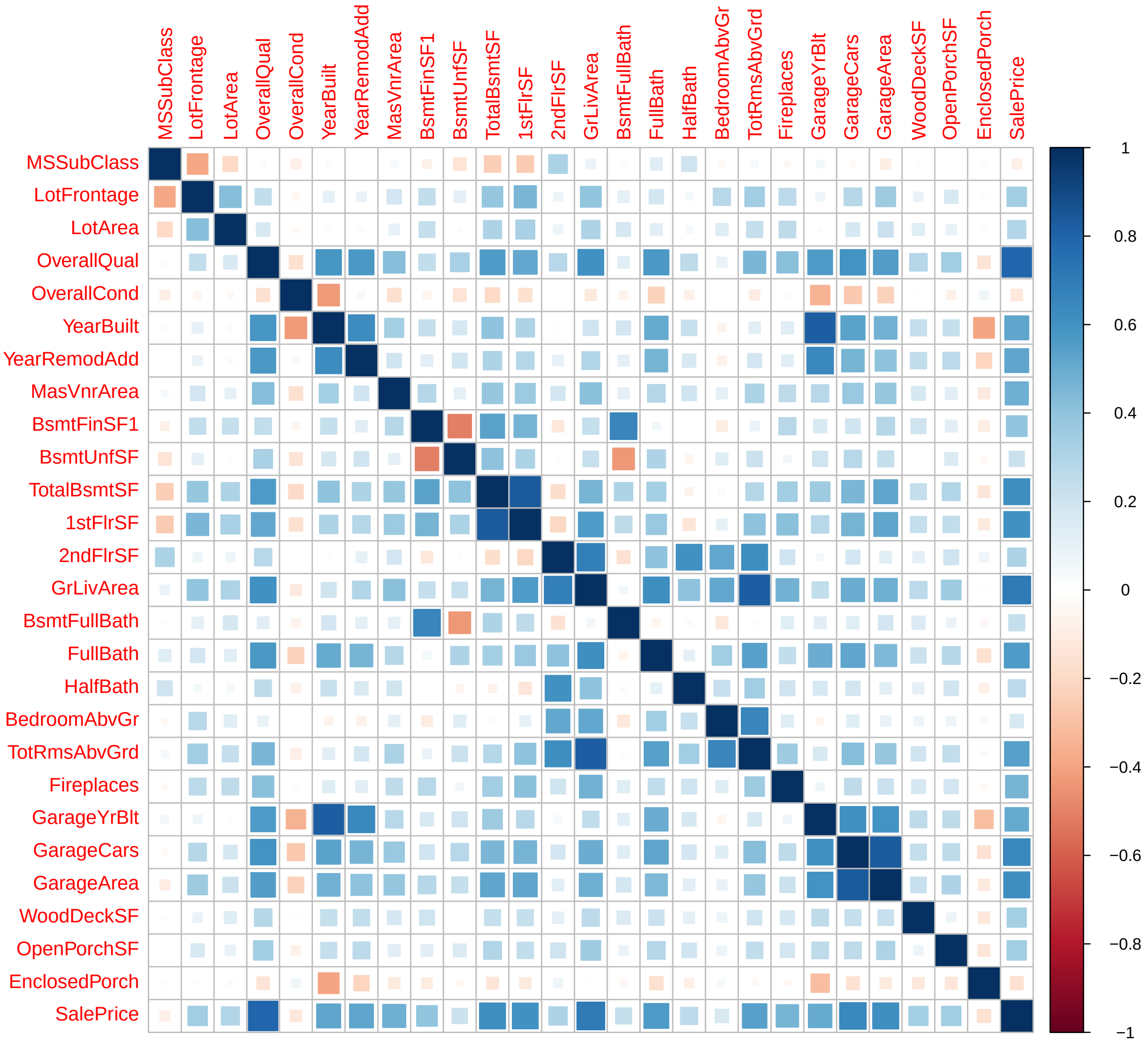
The histograms below show that majority of the houses have 2 full baths, 0 half baths, and have an average of 3 bedrooms.

```r
doPlots(train_cont, fun = plotHist, ii = 18:23, ncol = 2)
```



Explore the correlation

```
correlations <- cor(na.omit(train_cont[,-1, with = FALSE]))

# correlations
row_indic <- apply(correlations, 1, function(x) sum(x > 0.3 | x < -0.3) > 1)

correlations<- correlations[row_indic ,row_indic ]
corrplot(correlations, method="square")
```



Plot scatter plot for variables that have high correlation.

The correlation matrix below shows that there are several variables that are strongly and positively correlated with housing price.

High positive correlation:

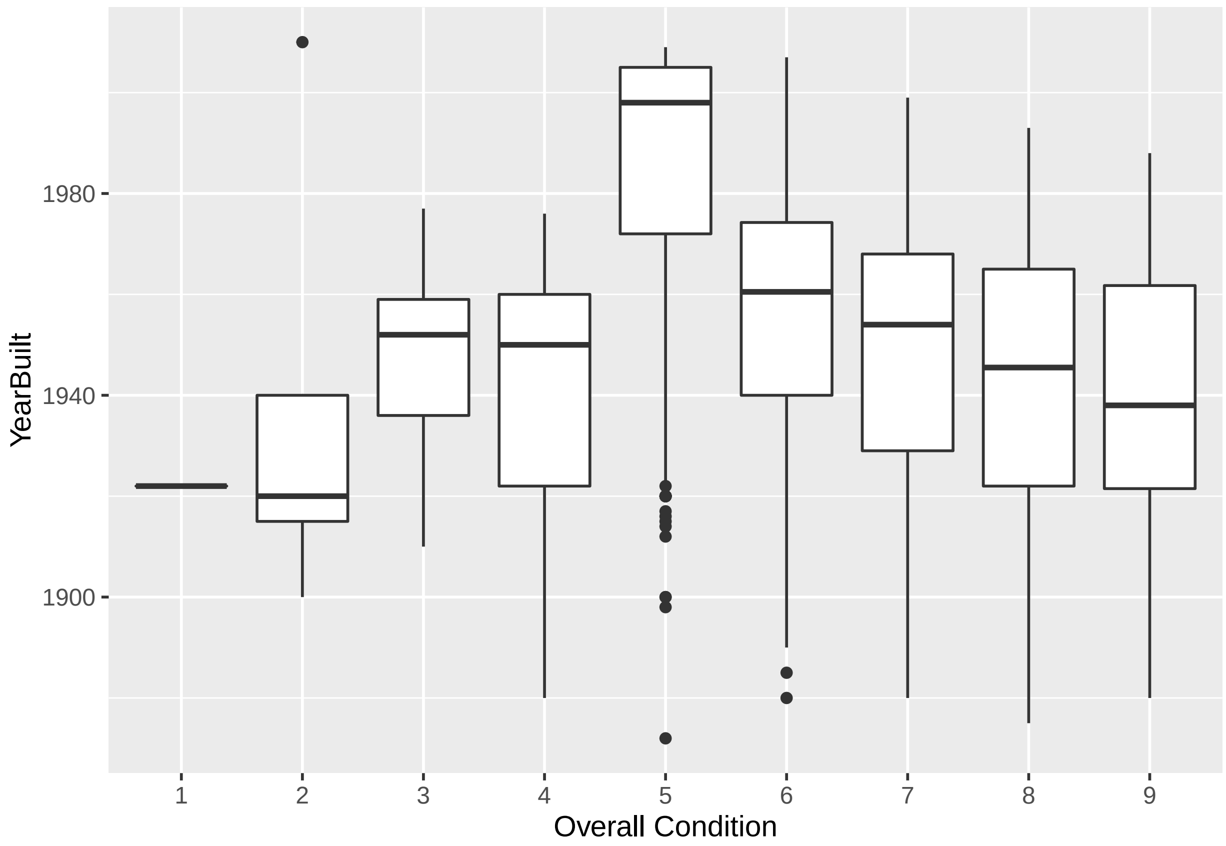
- OverallQual
- YearBuilt
- YearRemodAdd
- MasvnrArea
- BsmtFinSF1
- TotalBsmtSF

- 1stFlrSF
- GrLiveArea
- FullBath
- TotRmsAbvGrd
- FirePlaces
- GarageYrBlt
- GarageCars
- GarageArea
- WoodDeskSF
- OpenPorchSF

The number of enclosed porches are negatively correlated with year built. It seems that potential housebuyers do not want an enclosed porch and house developers have been building less enclosed porches in recent years. It is also negatively correlated with SalePrice, which makes sense.

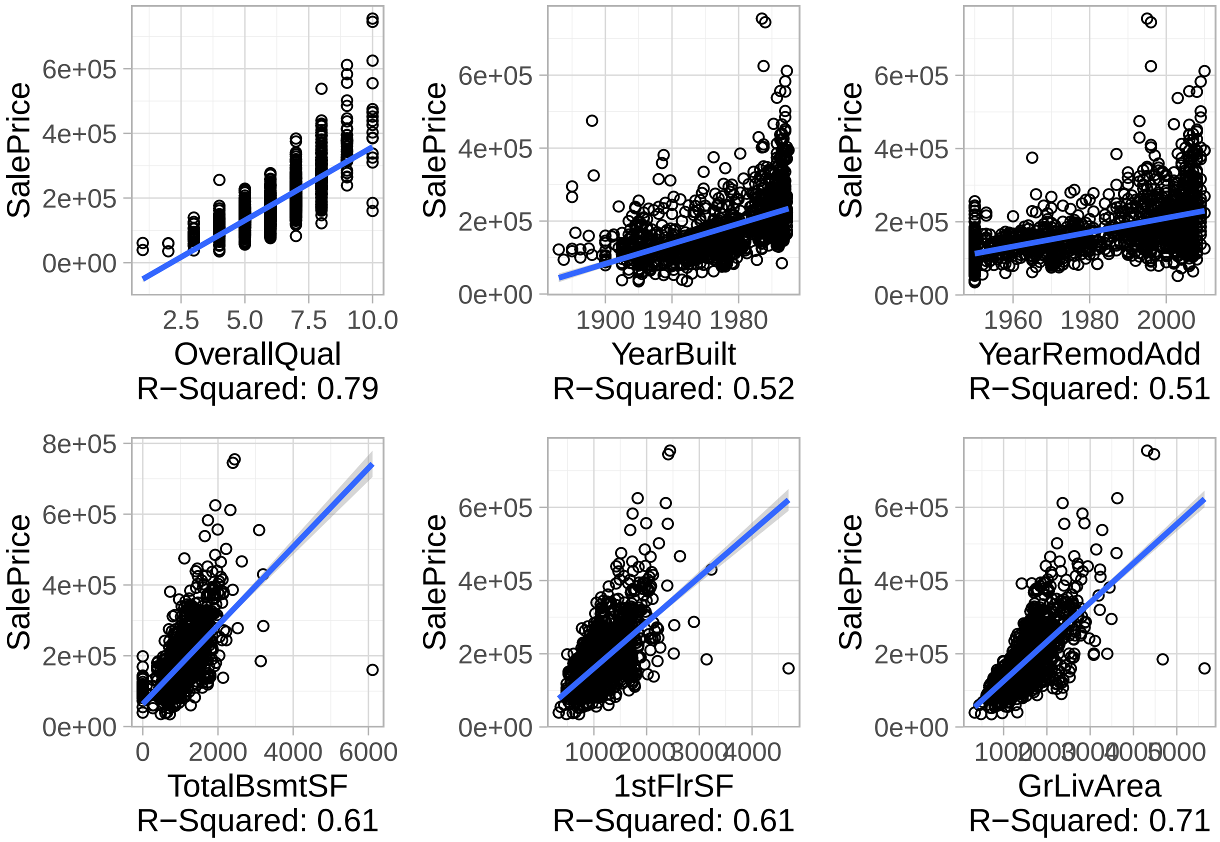
There is some slight negative correlation between OverallCond and SalePrice. There is also strong negative correlation between Yearbuilt and OverallCond. It seems to be that recently built houses tend to been in worse Overall Condition.

```
train %>% select(OverallCond, YearBuilt) %>% ggplot(aes(factor(OverallCond),YearBuilt)) + geom_boxplot(
```



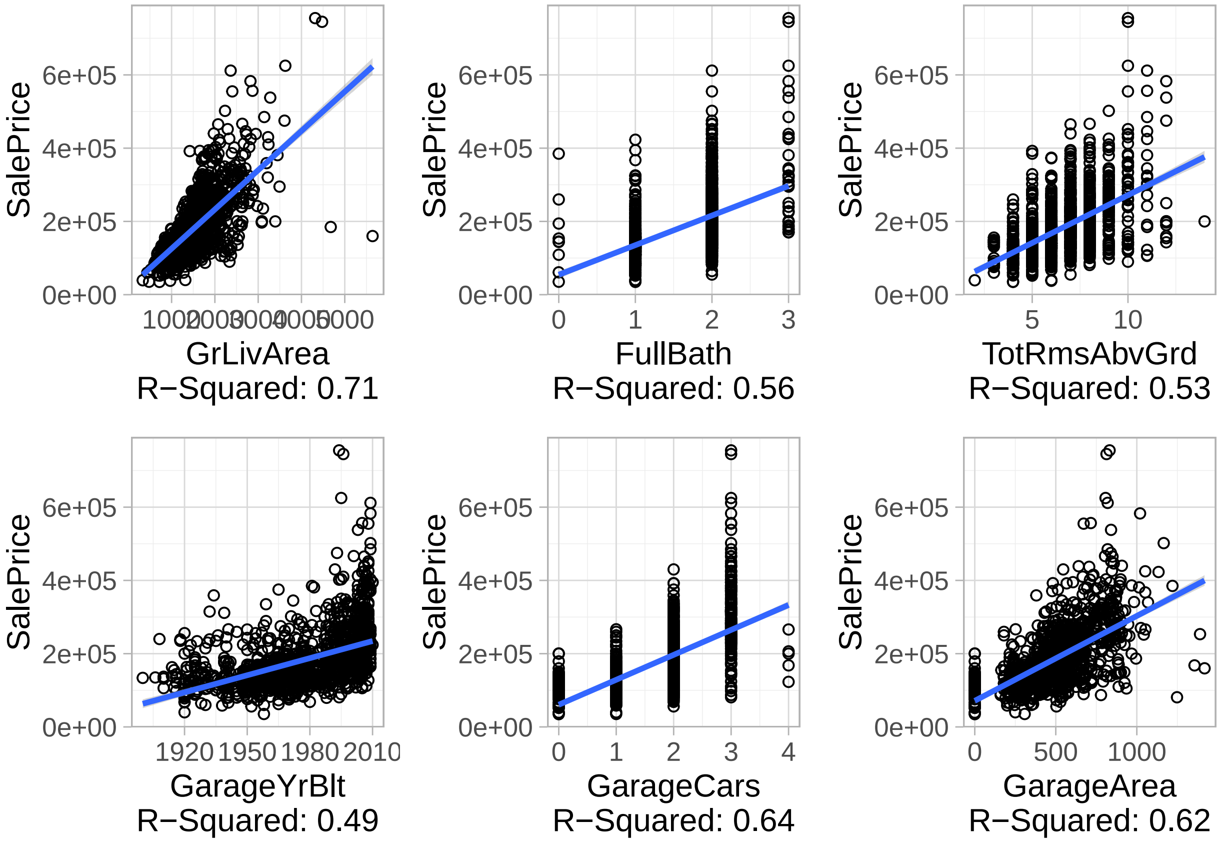
```
plotCorr <- function(data_in, i){
  data <- data.frame(x = data_in[[i]], SalePrice = data_in$SalePrice)
  p <- ggplot(data, aes(x = x, y = SalePrice)) + geom_point(shape = 1, na.rm = TRUE) + geom_smooth(meth
  return(suppressWarnings(p))
}
```

```
highcorr <- c(names(correlations[,'SalePrice'])[which(correlations[,'SalePrice'] > 0.5)], names(correla

data_corr <- train[,highcorr, with = FALSE]


doPlots(data_corr, fun = plotCorr, ii = 1:6)
```



```
doPlots(data_corr, fun = plotCorr, ii = 6:11)
```

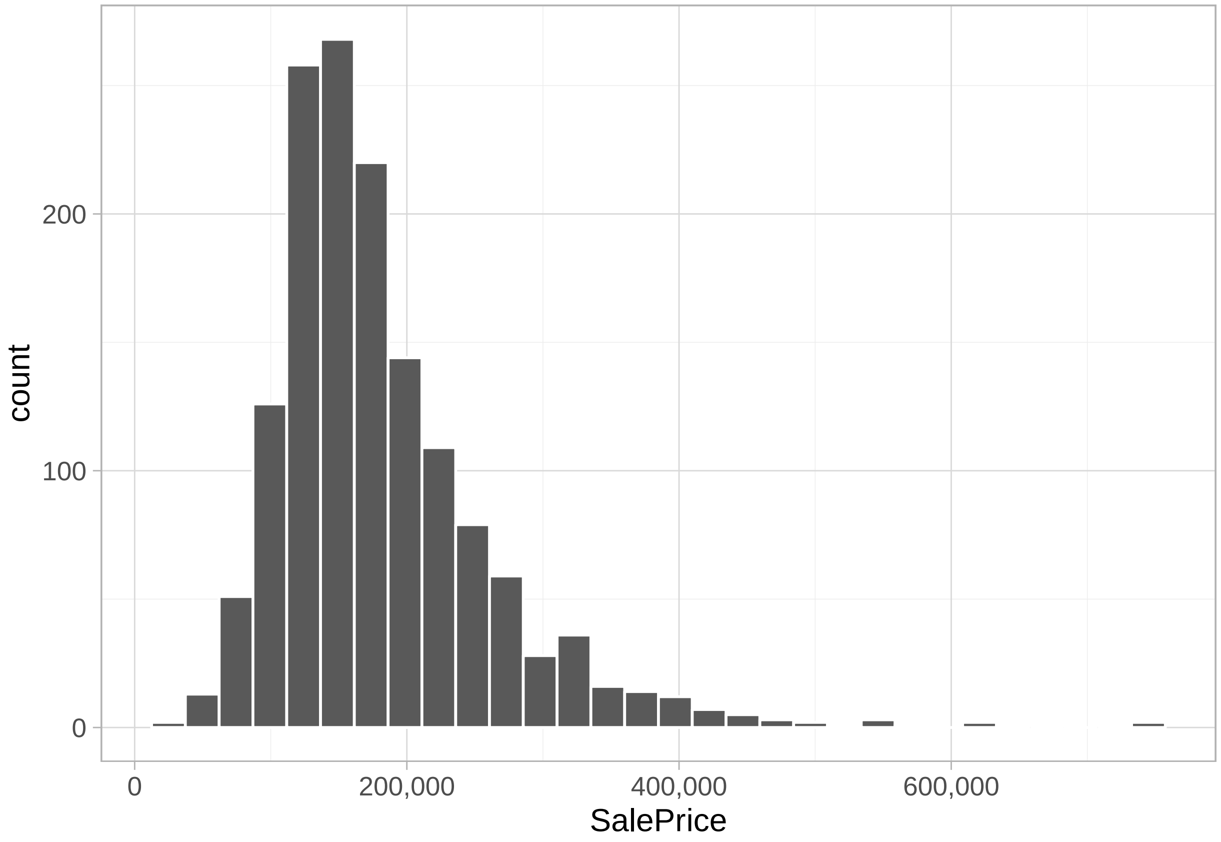
```
## Warning: Removed 81 rows containing non-finite values (stat_smooth).
```



The histogram for the response variable SalePrice shows that it is skewed. Taking the log of the variable normalizes it.

```
library(scales)
ggplot(train, aes(x=SalePrice)) + geom_histogram(col = 'white') + theme_light() +scale_x_continuous(labe
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

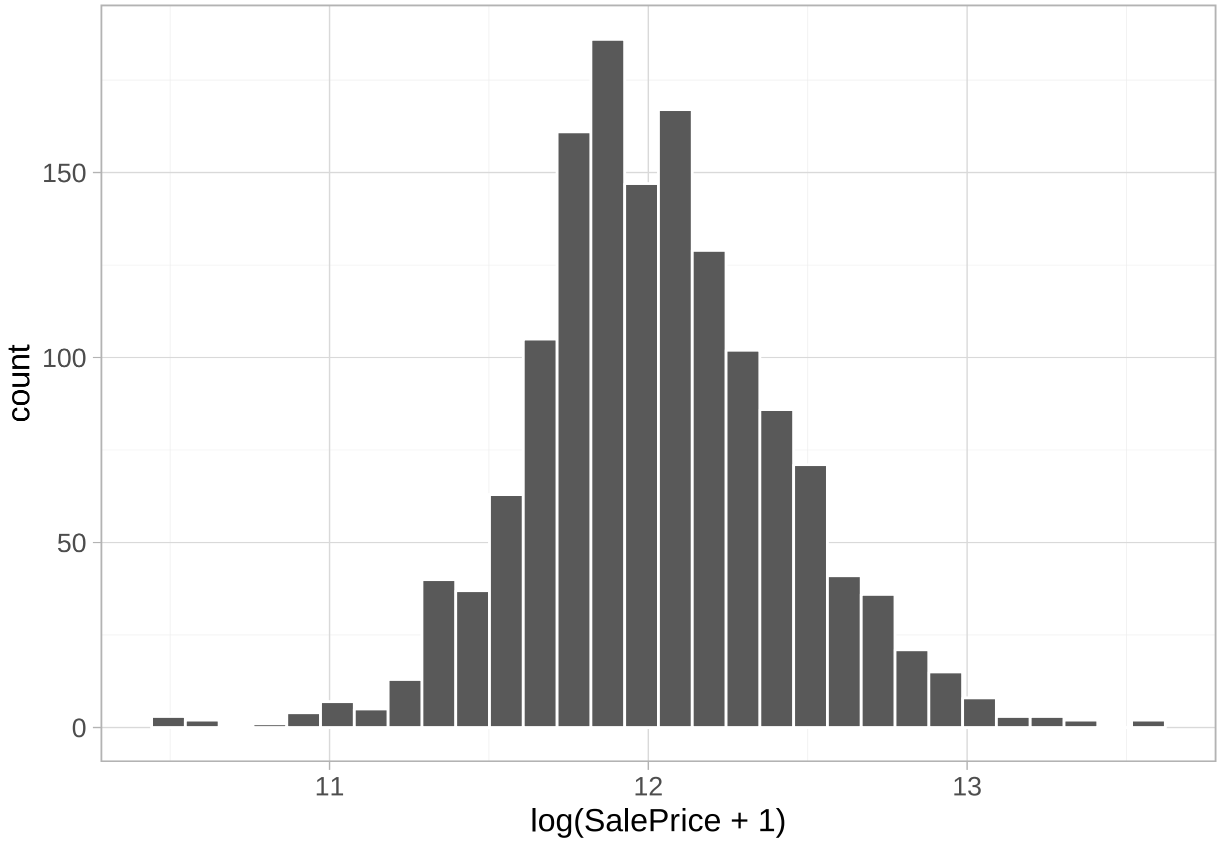


```r
summary(train[,.(SalePrice)])
```

```
##    SalePrice
##  Min.   : 34900
##  1st Qu.:129975
##  Median :163000
##  Mean   :180921
##  3rd Qu.:214000
##  Max.   :755000
```

```r
#Normalize distribution
ggplot(train, aes(x=log(SalePrice+1))) + geom_histogram(col = 'white') + theme_light()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


150
100
50
0
count
11
12
13
log(SalePrice + 1)