565 Final Project House Price EDA

Yu Chen, Zifei Dong, Ning Pan, Sifan Tao, Yao Yao

2023-03-07

House Prices Prediction

We take will use the House Prices (https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data) data from Kaggle. The measured variables can be seen here (https://github.com/Agaresd47/House_Price_Prediction/blob/main/house-prices-advanced-regression-techniques/Data%20Description.pdf)

Exploratory data analysis：

```
# Read the train.csv file into a data frame
house_prices <- read.csv("train.csv", header = TRUE, stringsAsFactors = FALSE)

dim(house_prices)
```

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

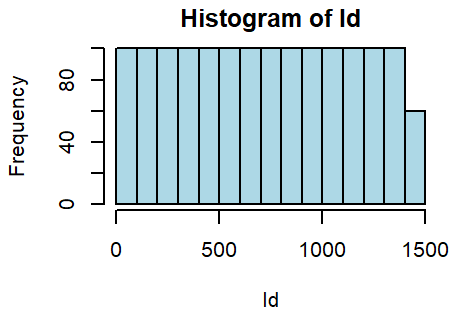
```
head(house_prices)
```

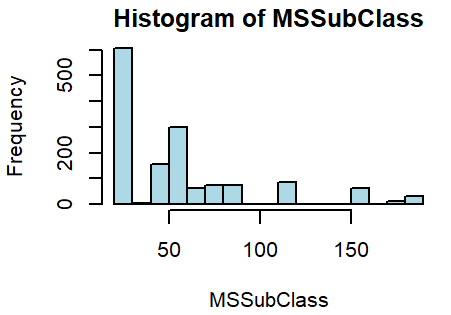
```
##   Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 1  1         60       RL          65    8450   Pave  <NA>      Reg         Lvl
## 2  2         20       RL          80    9600   Pave  <NA>      Reg         Lvl
## 3  3         60       RL          68   11250   Pave  <NA>      IR1         Lvl
## 4  4         70       RL          60    9550   Pave  <NA>      IR1         Lvl
## 5  5         60       RL          84   14260   Pave  <NA>      IR1         Lvl
## 6  6         50       RL          85   14115   Pave  <NA>      IR1         Lvl
##   Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
## 1    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 2    AllPub       FR2       Gtl      Veenker      Feedr       Norm     1Fam
## 3    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 4    AllPub    Corner       Gtl      Crawfor       Norm       Norm     1Fam
## 5    AllPub       FR2       Gtl      NoRidge       Norm       Norm     1Fam
## 6    AllPub    Inside       Gtl      Mitchel       Norm       Norm     1Fam
##   HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl
## 1     2Story           7           5      2003         2003     Gable  CompShg
## 2     1Story           6           8      1976         1976     Gable  CompShg
## 3     2Story           7           5      2001         2002     Gable  CompShg
## 4     2Story           7           5      1915         1970     Gable  CompShg
## 5     2Story           8           5      2000         2000     Gable  CompShg
## 6     1.5Fin           5           5      1993         1995     Gable  CompShg
##   Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation
## 1     VinylSd     VinylSd    BrkFace        196        Gd        TA      PConc
## 2     MetalSd     MetalSd       None          0        TA        TA     CBlock
## 3     VinylSd     VinylSd    BrkFace        162        Gd        TA      PConc
## 4     Wd Sdng     Wd Shng       None          0        TA        TA     BrkTil
## 5     VinylSd     VinylSd    BrkFace        350        Gd        TA      PConc
## 6     VinylSd     VinylSd       None          0        TA        TA       Wood
##   BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
## 1       Gd       TA           No          GLQ        706          Unf
## 2       Gd       TA           Gd          ALQ        978          Unf
## 3       Gd       TA           Mn          GLQ        486          Unf
## 4       TA       Gd           No          ALQ        216          Unf
## 5       Gd       TA           Av          GLQ        655          Unf
## 6       Gd       TA           No          GLQ        732          Unf
##   BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical
## 1          0       150         856    GasA        Ex          Y      SBrkr
## 2          0       284        1262    GasA        Ex          Y      SBrkr
## 3          0       434         920    GasA        Ex          Y      SBrkr
## 4          0       540         756    GasA        Gd          Y      SBrkr
## 5          0       490        1145    GasA        Ex          Y      SBrkr
## 6          0        64         796    GasA        Ex          Y      SBrkr
##   X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
## 1       856       854            0      1710            1            0        2
## 2      1262         0            0      1262            0            1        2
## 3       920       866            0      1786            1            0        2
## 4       961       756            0      1717            1            0        1
## 5      1145      1053            0      2198            1            0        2
## 6       796       566            0      1362            1            0        1
##   HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
## 1        1            3            1          Gd            8        Typ
## 2        0            3            1          TA            6        Typ
## 3        1            3            1          Gd            6        Typ
## 4        0            3            1          Gd            7        Typ
## 5        1            4            1          Gd            9        Typ
```

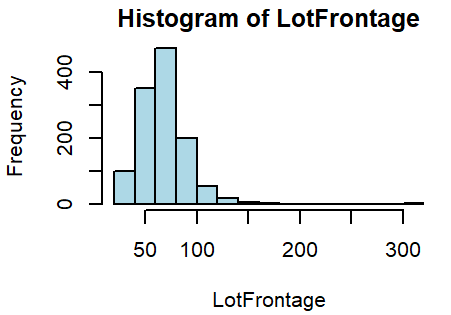
```
## 6        1            1            1          TA            5        Typ
##   Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
## 1          0       <NA>     Attchd        2003          RFn          2
## 2          1         TA     Attchd        1976          RFn          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 X3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature
## 1             0          0           0        0   <NA>  <NA>        <NA>
## 2             0          0           0        0   <NA>  <NA>        <NA>
## 3             0          0           0        0   <NA>  <NA>        <NA>
## 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
```

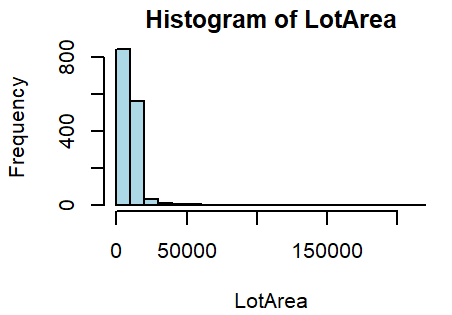
Distributions of prostate cancer variables

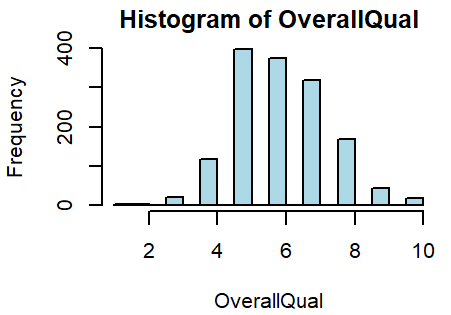
```
# Get a vector of variable names that are numeric
numeric_vars <- names(house_prices)[sapply(house_prices, is.numeric)]

# Setup grid and margins for plotting
par(mfrow=c(3, 3), mar=c(4, 4, 2, 0.5))

# Loop through each numeric variable and plot its distribution
for (j in 1:length(numeric_vars)) {
  hist(house_prices[, numeric_vars[j]], xlab=numeric_vars[j],
       main=paste("Histogram of", numeric_vars[j]),
       col="lightblue", breaks=20)
}
```

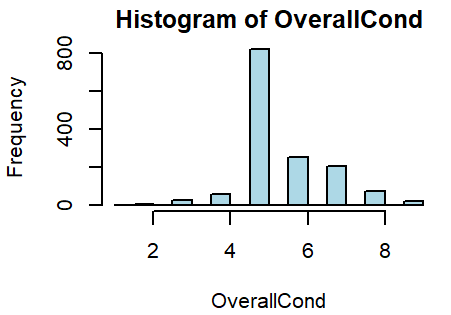

Histogram of Id
Frequency
0
40
80
0
500
1000
1500
Id

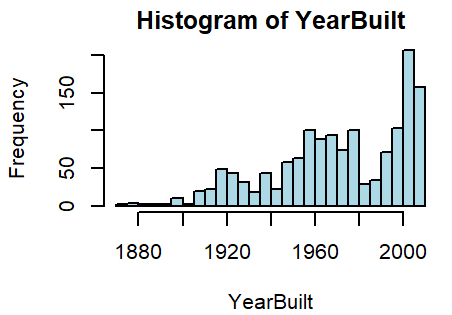

Histogram of MSSubClass
Frequency
0
200
500
50
100
150
MSSubClass

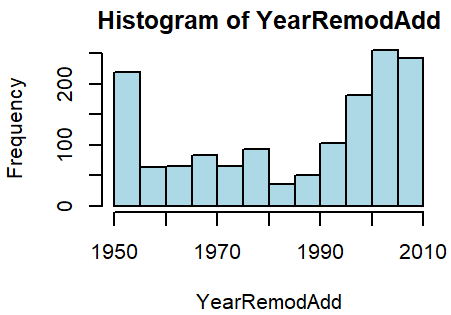

Histogram of LotFrontage
Frequency
0
200
400
50
100
200
300
LotFrontage

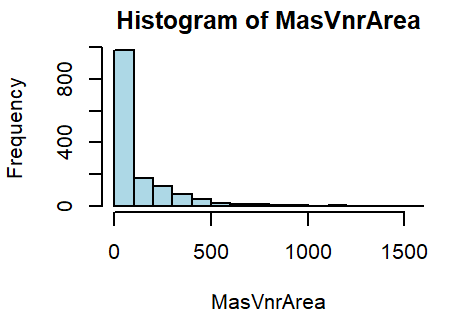

Histogram of LotArea
Frequency
0
400
800
0
50000
150000
LotArea

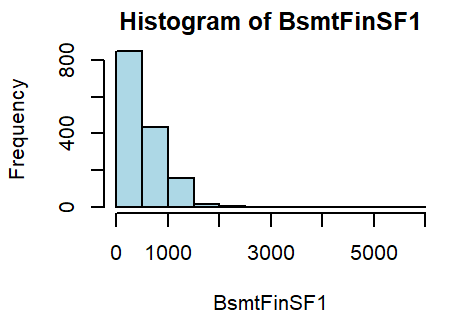

Histogram of OverallQual
Frequency
0
200
400
2
4
6
8
10
OverallQual

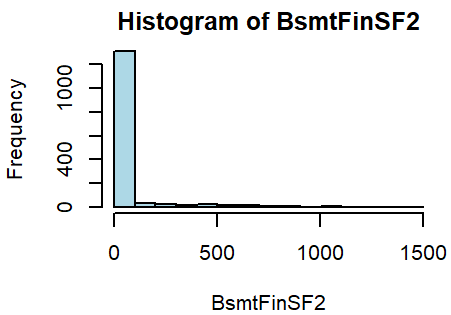

Histogram of OverallCond
Frequency
0
400
800
2
4
6
8
OverallCond

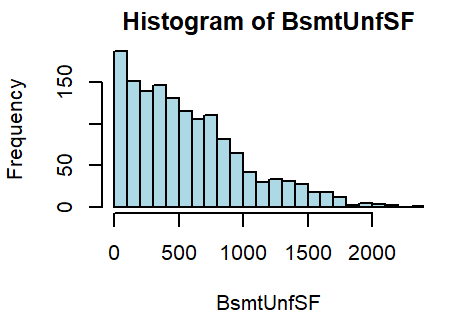

Histogram of YearBuilt
Frequency
0
50
150
1880
1920
1960
2000
YearBuilt

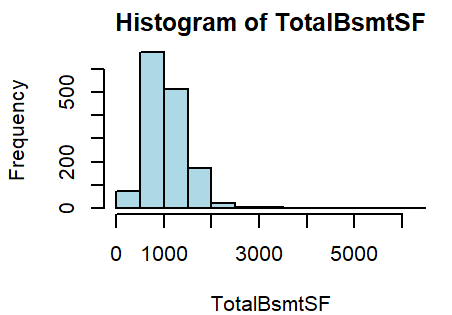

Histogram of YearRemodAdd
Frequency
0
100
200
1950
1970
1990
2010
YearRemodAdd

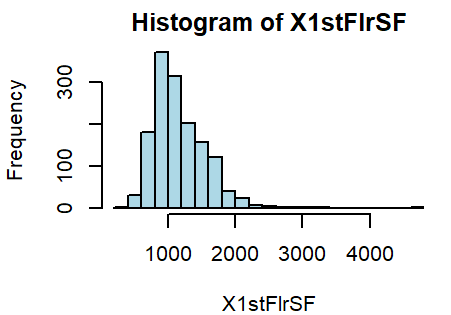

Histogram of MasVnrArea
Frequency
0
400
800
0
500
1000
1500
MasVnrArea

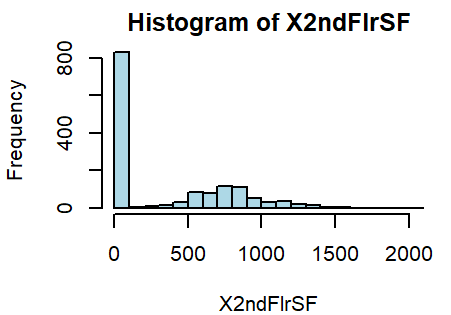

Histogram of BsmtFinSF1
Frequency
0
400
800
0
1000
3000
5000
BsmtFinSF1

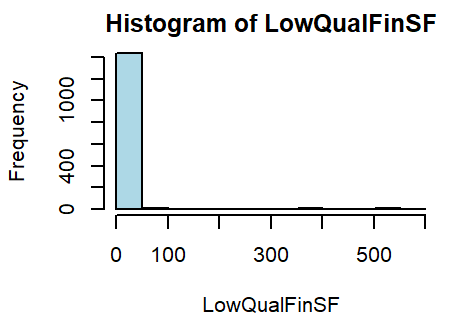

Histogram of BsmtFinSF2
Frequency
0
400
1000
0
500
1000
1500
BsmtFinSF2

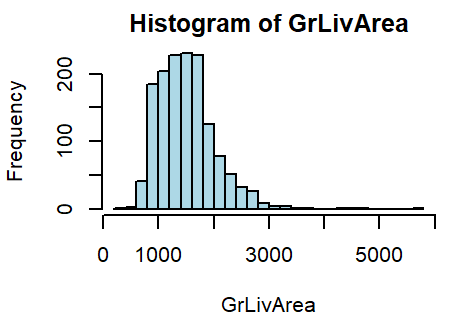

Histogram of BsmtUnfSF
Frequency
0
50
150
0
500
1000
1500
2000
BsmtUnfSF

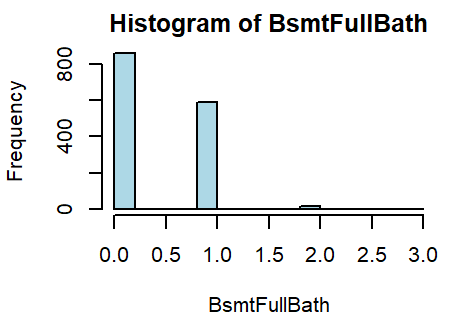

Histogram of TotalBsmtSF
Frequency
0
200
500
0
1000
3000
5000
TotalBsmtSF

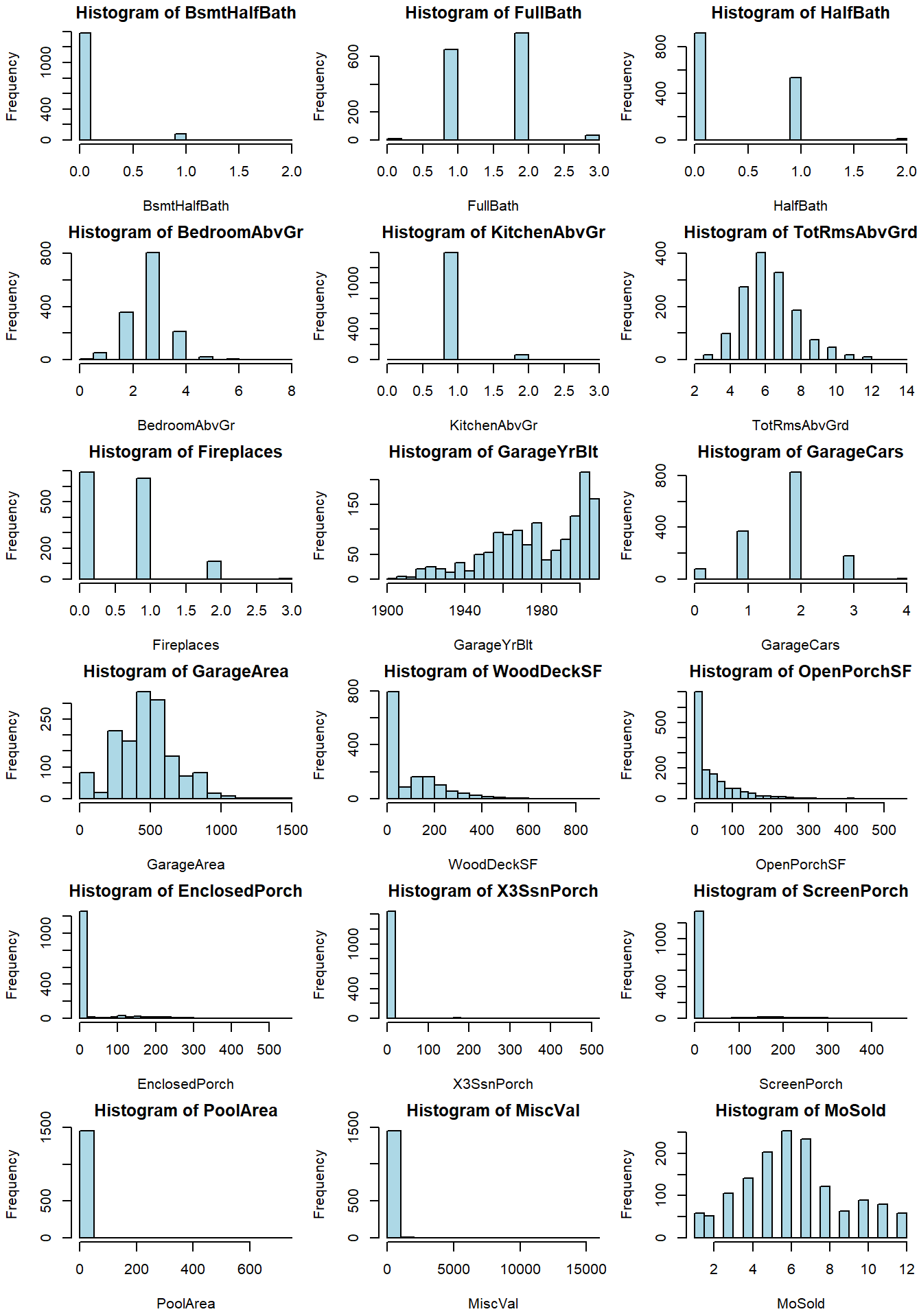

Histogram of X1stFlrSF
Frequency
0
100
300
1000
2000
3000
4000
X1stFlrSF

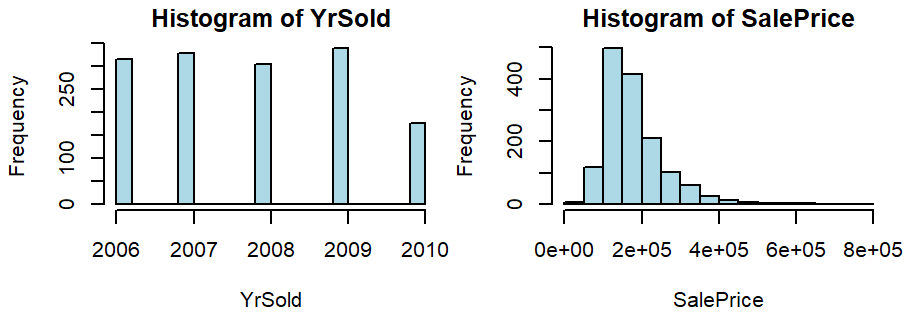

Histogram of X2ndFlrSF
Frequency
0
400
800
0
500
1000
1500
2000
X2ndFlrSF


Histogram of LowQualFinSF
Frequency
0
400
1000
0
100
300
500
LowQualFinSF


Histogram of GrLivArea
Frequency
0
100
200
0
1000
3000
5000
GrLivArea


Histogram of BsmtFullBath
Frequency
0
400
800
0.0
0.5
1.0
1.5
2.0
2.5
3.0
BsmtFullBath

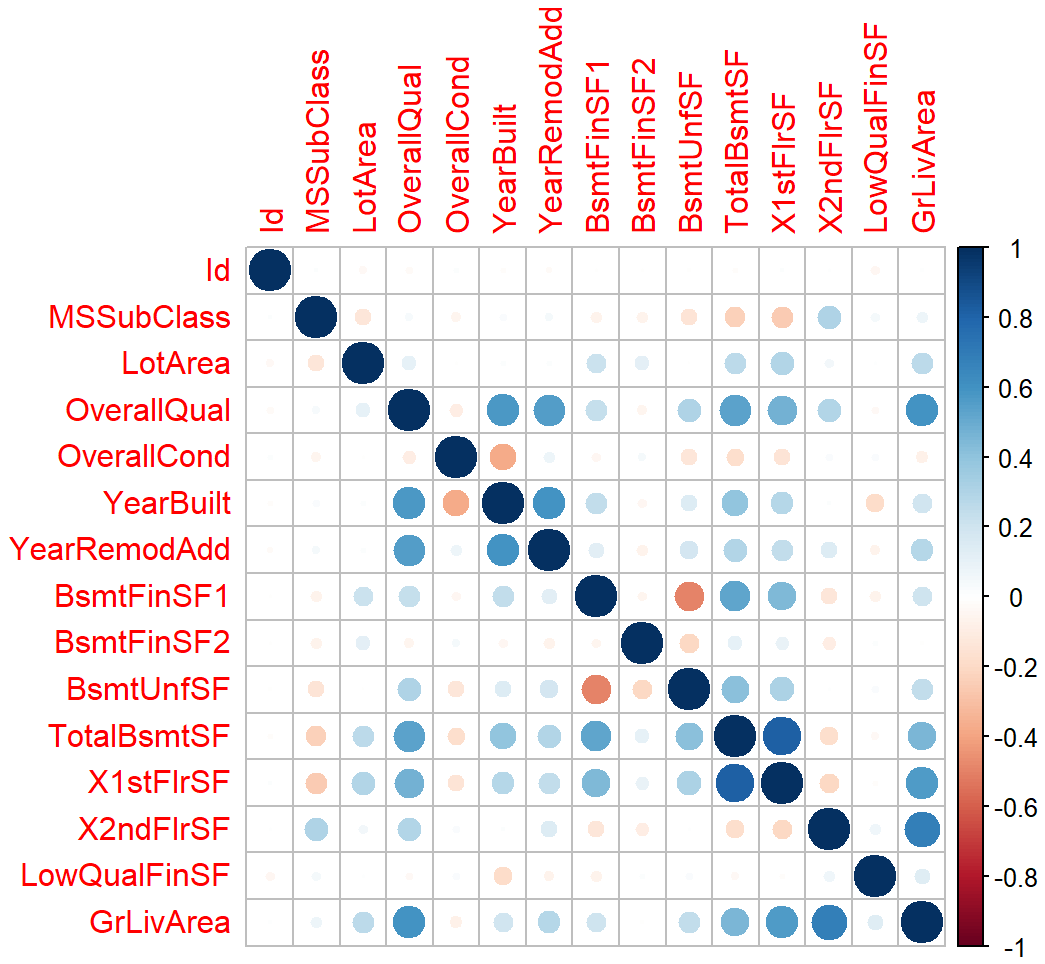

Histogram of BsmtHalfBath
Frequency
BsmtHalfBath
Histogram of FullBath
Frequency
FullBath
Histogram of HalfBath
Frequency
HalfBath
Histogram of BedroomAbvGr
Frequency
BedroomAbvGr
Histogram of KitchenAbvGr
Frequency
KitchenAbvGr
Histogram of TotRmsAbvGrd
Frequency
TotRmsAbvGrd
Histogram of Fireplaces
Frequency
Fireplaces
Histogram of GarageYrBlt
Frequency
GarageYrBlt
Histogram of GarageCars
Frequency
GarageCars
Histogram of GarageArea
Frequency
GarageArea
Histogram of WoodDeckSF
Frequency
WoodDeckSF
Histogram of OpenPorchSF
Frequency
OpenPorchSF
Histogram of EnclosedPorch
Frequency
EnclosedPorch
Histogram of X3SsnPorch
Frequency
X3SsnPorch
Histogram of ScreenPorch
Frequency
ScreenPorch
Histogram of PoolArea
Frequency
PoolArea
Histogram of MiscVal
Frequency
MiscVal
Histogram of MoSold
Frequency
MoSold



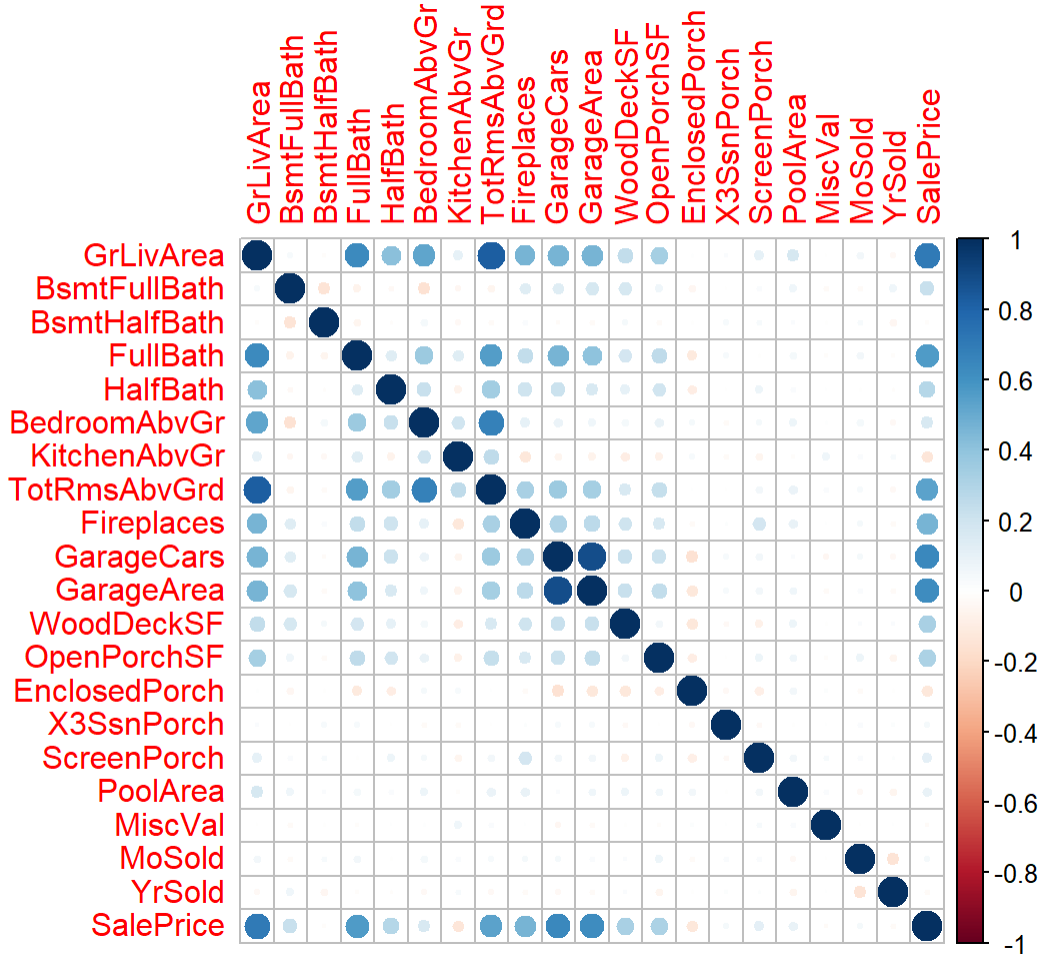
It is apparent that the majority of data has a strong skew and pattern.

Correlations between House Price variables

```
corrplot(cor(house_prices_new[c(1:15)]), type = "full")
```




```
corrplot(cor(house_prices_new[c(15:35)]), type = "full")
```



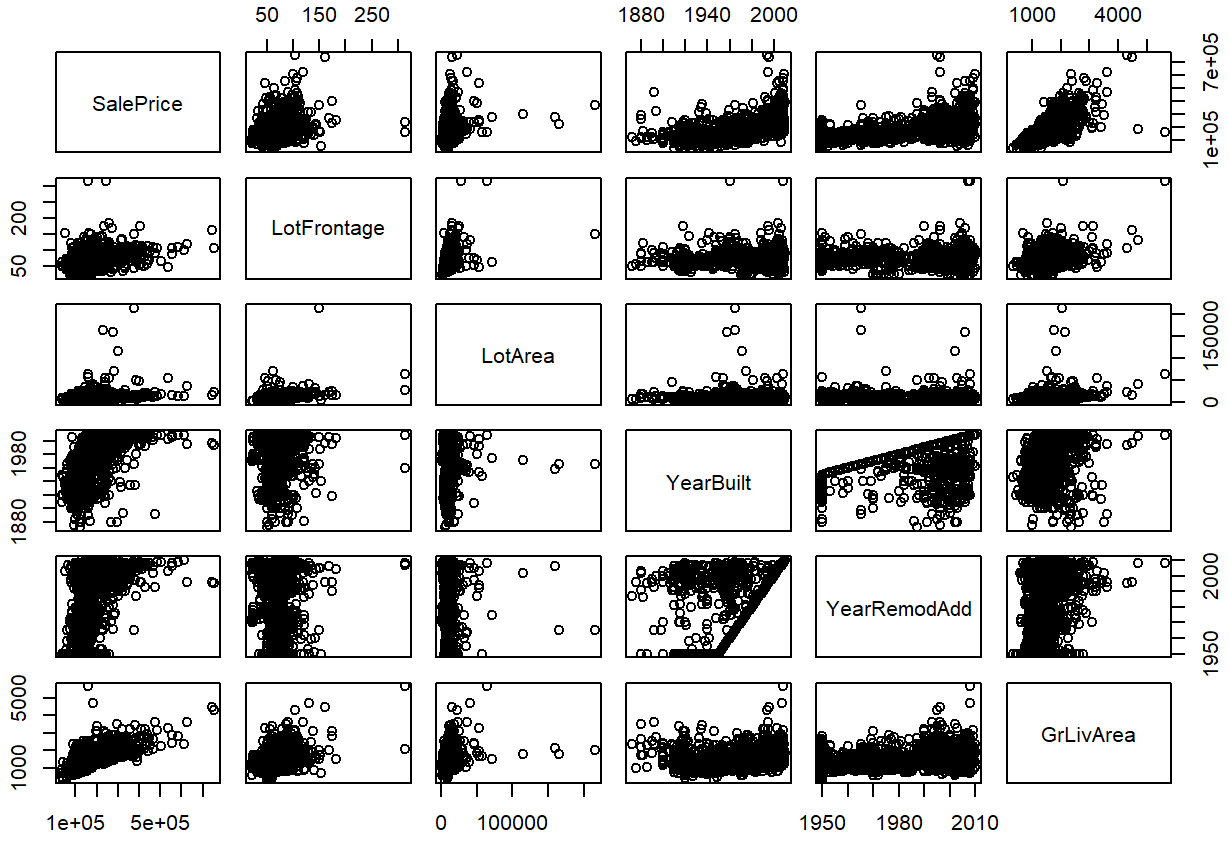
The majority of variables have a moderate relationship between each other.

Visualizing Relationships Among Variables

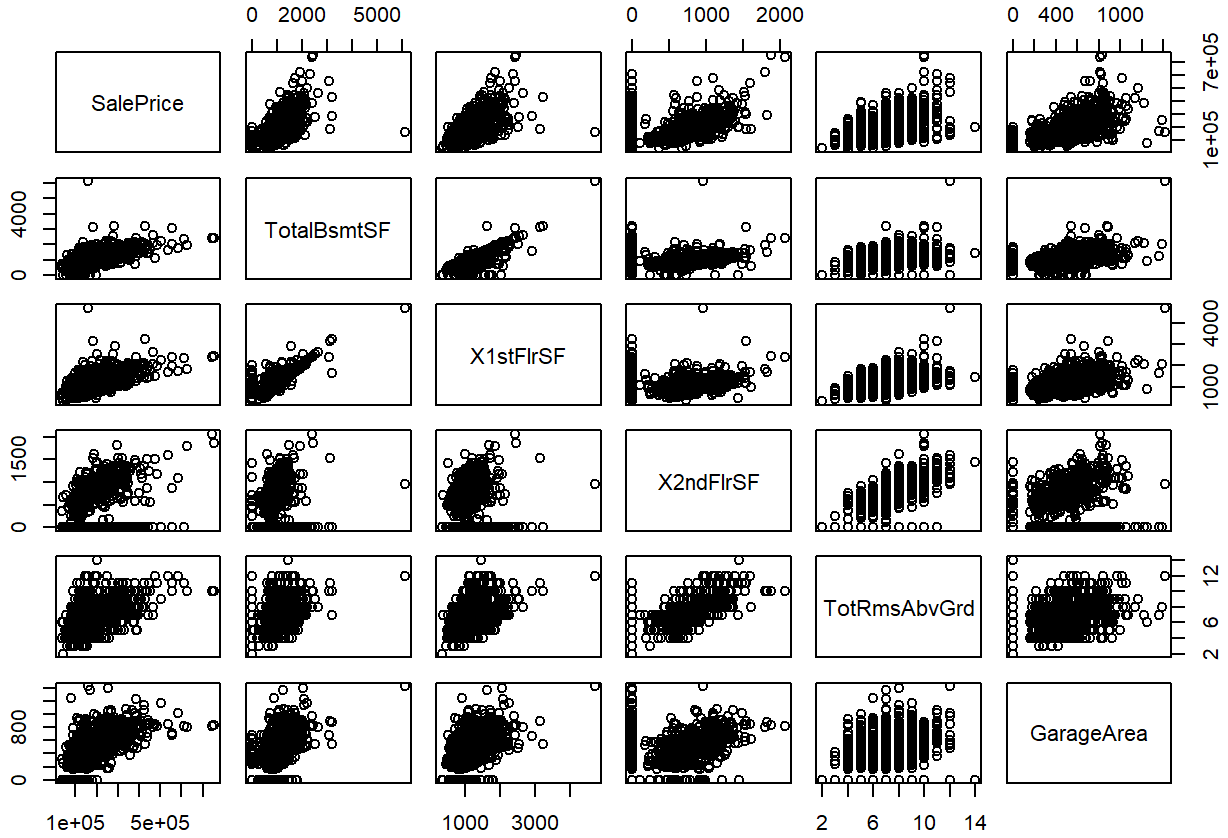
```
names(house_prices)[sapply(house_prices, is.numeric)]
```

```
##  [1] "Id"            "MSSubClass"    "LotFrontage"   "LotArea"
##  [5] "OverallQual"   "OverallCond"   "YearBuilt"     "YearRemodAdd"
##  [9] "MasVnrArea"    "BsmtFinSF1"    "BsmtFinSF2"    "BsmtUnfSF"
## [13] "TotalBsmtSF"   "X1stFlrSF"     "X2ndFlrSF"     "LowQualFinSF"
## [17] "GrLivArea"     "BsmtFullBath"  "BsmtHalfBath"  "FullBath"
## [21] "HalfBath"      "BedroomAbvGr"  "KitchenAbvGr"  "TotRmsAbvGrd"
## [25] "Fireplaces"    "GarageYrBlt"   "GarageCars"    "GarageArea"
## [29] "WoodDeckSF"    "OpenPorchSF"   "EnclosedPorch" "X3SsnPorch"
## [33] "ScreenPorch"   "PoolArea"      "MiscVal"       "MoSold"
## [37] "YrSold"        "SalePrice"
```

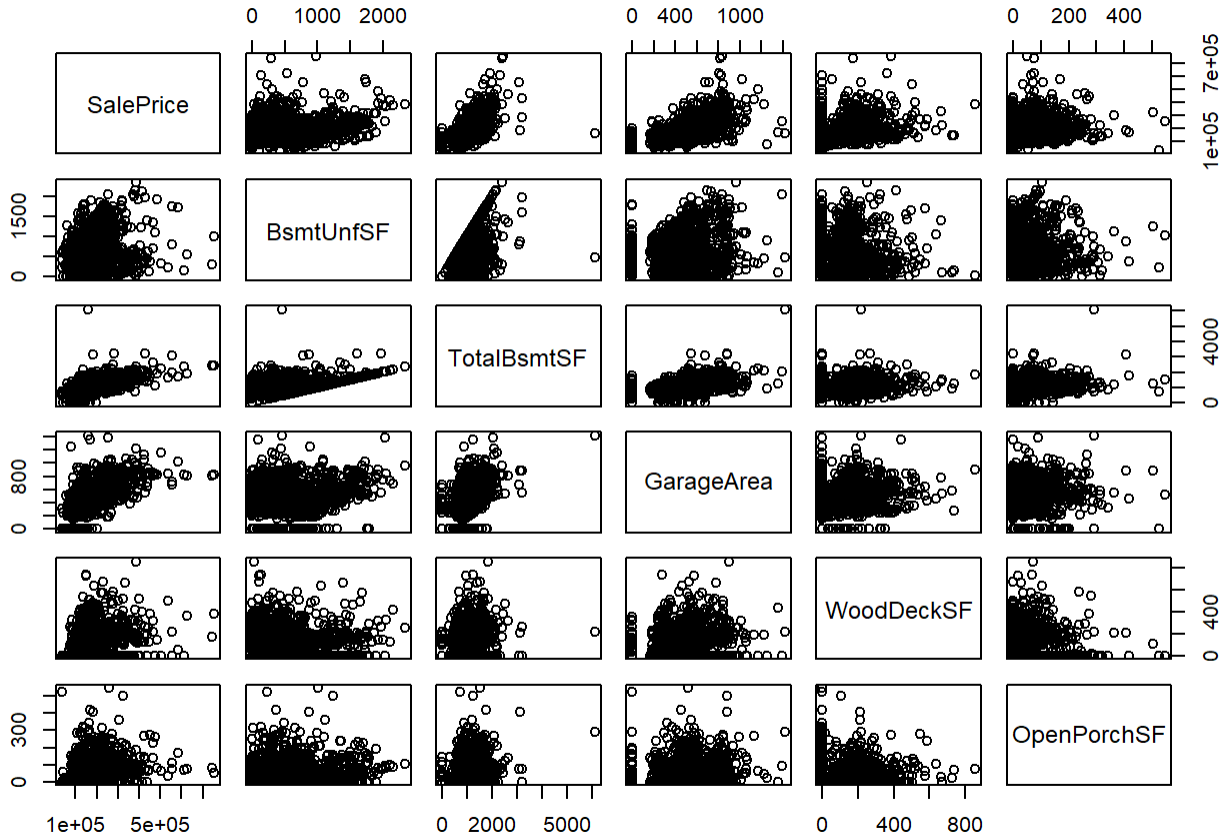
```
#Sale price and General housing information
pairs(~ SalePrice+LotFrontage+LotArea+YearBuilt+YearRemodAdd+GrLivArea, data=house_prices)
```




```
#Sale price and Major surface areas
pairs(~ SalePrice+TotalBsmtSF+X1stFlrSF+X2ndFlrSF+TotRmsAbvGrd+GarageArea, data=house_prices)
```




```
#Sale price and Basement and "Add-on" areas
pairs(~ SalePrice+BsmtUnfSF+TotalBsmtSF+GarageArea+WoodDeckSF+OpenPorchSF, data=house_prices)
```



We excluded Variables due to their insignificant information and fixed scales that graphs cannot represent:

OverallQual, OverallCond, BsmtFinSF1, BsmtFinSF2, MSSubClass, LowQualFinSF, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, BedroomAbvGr, KitchenAbvGr, Fireplaces, GarageYrBlt, GarageCars, EnclosedPorch, X3SsnPorch, ScreenPorch, PoolArea, MiscVal, MoSold, YrSold