Project_HousePrediction_Initial Data Exploration

Hyunkyung Kim

October 11, 2018

Import Data and Save a Copy

housedata - original data housedatac - working copy of the data

```
housedata<-read.csv("C:\\Users\\Hyunkyung
Kim\\Desktop\\CKME999\\136\\dataset\\all\\train.csv")

housedatac<-housedata
```

Investigating the variables.

```
str(housedata)
```

```
## '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     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5
4 ...
##  $ 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       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2
...
##  $ Alley        : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA
NA NA NA ...
##  $ 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   : int  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/ 4 levels "Ex","Fa","Gd",..: 3 3 3 4 3 3 1 3 4
4 ...
##  $ BsmtCond     : Factor w/ 4 levels "Fa","Gd","Po",..: 4 4 4 2 4 4 4 4 4
4 ...
##  $ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4
4 ...
##  $ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1
6 3 ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 6 6 6 6 6 6 6 2
6 6 ...
##  $ 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/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2
5 ...
##  $ 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/ 5 levels "Ex","Fa","Gd",..: NA 5 5 3 5 NA 3 5
5 5 ...
##  $ GarageType   : Factor w/ 6 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2
6 2 ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939
...
##  $ GarageFinish : Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3
2 ...
##  $ 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/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 2
3 ...
##  $ GarageCond   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5
5 ...
##  $ 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/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA
NA NA NA ...
##  $ Fence        : Factor w/ 4 levels "GdPrv","GdWo",..: NA NA NA NA NA 3
NA NA NA NA ...
##  $ MiscFeature  : Factor w/ 4 levels "Gar2","Othr",..: NA NA NA NA NA 3 NA
3 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     : 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 ...
```

Checking for duplicates & uniques : Both 1460 rows - meaning No duplicate

Display features that has non zero N/As.

```
nrow(housedata[,-1])
```

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

```
nrow(unique(housedata[,-1]))
```

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

```
NAs<-colSums(is.na(housedata))
NAs[NAs>0]
```

```
##  LotFrontage        Alley   MasVnrType   MasVnrArea     BsmtQual
##          259         1369            8            8           37
##     BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2   Electrical
##           37           38           37           38            1
##  FireplaceQu   GarageType  GarageYrBlt GarageFinish   GarageQual
##          690           81           81           81           81
##   GarageCond       PoolQC        Fence  MiscFeature
##           81         1453         1179         1406
```

Number of rows = # of unique rows : no duplicate data Features that has N/As are displayed - Will work with that in the next step.

Data Description

Checking Min/Max/quartiles for numerical, and table for categorical variables.

```
summary(housedata)
```

```
##        Id           MSSubClass       MSZoning     LotFrontage
##  Min.   :   1.0   Min.   : 20.0   C (all):  10   Min.   : 21.00
##  1st Qu.: 365.8   1st Qu.: 20.0   FV     :  65   1st Qu.: 59.00
##  Median : 730.5   Median : 50.0   RH     :  16   Median : 69.00
##  Mean   : 730.5   Mean   : 56.9   RL     :1151   Mean   : 70.05
##  3rd Qu.:1095.2   3rd Qu.: 70.0   RM     : 218   3rd Qu.: 80.00
##  Max.   :1460.0   Max.   :190.0                   Max.   :313.00
##                                                  NA's   :259
##     LotArea        Street      Alley      LotShape  LandContour
##  Min.   :  1300   Grvl:   6   Grvl:  50   IR1:484   Bnk:  63
##  1st Qu.:  7554   Pave:1454   Pave:  41   IR2: 41   HLS:  50
##  Median :  9478               NA's:1369   IR3: 10   Low:  36
##  Mean   : 10517                           Reg:925   Lvl:1311
##  3rd Qu.: 11602
##  Max.   :215245
##
##   Utilities      LotConfig    LandSlope   Neighborhood   Condition1
##  AllPub:1459   Corner : 263   Gtl:1382   NAmes  :225   Norm   :1260
##  NoSeWa:   1   CulDSac:  94   Mod:  65   CollgCr:150   Feedr  :  81
##                FR2    :  47   Sev:  13   OldTown:113   Artery :  48
```

```
##                FR3    :   4              Edwards:100   RRAn   :  26
##                Inside :1052              Somerst: 86   PosN   :  19
##                                          Gilbert: 79   RRAe   :  11
##                                          (Other):707   (Other):  15
##    Condition2     BldgType      HouseStyle   OverallQual
##  Norm   :1445   1Fam  :1220   1Story :726   Min.   : 1.000
##  Feedr  :   6   2fmCon:  31   2Story :445   1st Qu.: 5.000
##  Artery :   2   Duplex:  52   1.5Fin :154   Median : 6.000
##  PosN   :   2   Twnhs :  43   SLvl   : 65   Mean   : 6.099
##  RRNn   :   2   TwnhsE: 114   SFoyer : 37   3rd Qu.: 7.000
##  PosA   :   1                 1.5Unf : 14   Max.   :10.000
##  (Other):   2                 (Other): 19
##   OverallCond      YearBuilt     YearRemodAdd    RoofStyle
##  Min.   :1.000   Min.   :1872   Min.   :1950   Flat   :  13
##  1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   Gable  :1141
##  Median :5.000   Median :1973   Median :1994   Gambrel:  11
##  Mean   :5.575   Mean   :1971   Mean   :1985   Hip    : 286
##  3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004   Mansard:   7
##  Max.   :9.000   Max.   :2010   Max.   :2010   Shed   :   2
##
##     RoofMatl     Exterior1st   Exterior2nd    MasVnrType    MasVnrArea
##  CompShg:1434   VinylSd:515   VinylSd:504   BrkCmn : 15   Min.   :   0.0
##  Tar&Grv:  11   HdBoard:222   MetalSd:214   BrkFace:445   1st Qu.:   0.0
##  WdShngl:   6   MetalSd:220   HdBoard:207   None   :864   Median :   0.0
##  WdShake:   5   Wd Sdng:206   Wd Sdng:197   Stone  :128   Mean   : 103.7
##  ClyTile:   1   Plywood:108   Plywood:142   NA's   :  8   3rd Qu.: 166.0
##  Membran:   1   CemntBd: 61   CmentBd: 60                 Max.   :1600.0
##  (Other):   2   (Other):128   (Other):136                 NA's   :8
##  ExterQual ExterCond  Foundation  BsmtQual   BsmtCond    BsmtExposure
##  Ex: 52    Ex:   3   BrkTil:146   Ex  :121   Fa  :  45   Av  :221
##  Fa: 14    Fa:  28   CBlock:634   Fa  : 35   Gd  :  65   Gd  :134
##  Gd:488    Gd: 146   PConc :647   Gd  :618   Po  :   2   Mn  :114
##  TA:906    Po:   1   Slab  : 24   TA  :649   TA  :1311   No  :953
##            TA:1282   Stone :  6   NA's: 37   NA's:  37   NA's: 38
##                      Wood  :  3
##
##  BsmtFinType1   BsmtFinSF1     BsmtFinType2   BsmtFinSF2
##  ALQ :220     Min.   :   0.0   ALQ :  19    Min.   :   0.00
##  BLQ :148     1st Qu.:   0.0   BLQ :  33    1st Qu.:   0.00
##  GLQ :418     Median : 383.5   GLQ :  14    Median :   0.00
##  LwQ : 74     Mean   : 443.6   LwQ :  46    Mean   :  46.55
##  Rec :133     3rd Qu.: 712.2   Rec :  54    3rd Qu.:   0.00
##  Unf :430     Max.   :5644.0   Unf :1256    Max.   :1474.00
##  NA's: 37                      NA's:  38
##    BsmtUnfSF       TotalBsmtSF      Heating     HeatingQC CentralAir
##  Min.   :   0.0   Min.   :   0.0   Floor:   1   Ex:741    N:  95
##  1st Qu.: 223.0   1st Qu.: 795.8   GasA :1428   Fa: 49    Y:1365
##  Median : 477.5   Median : 991.5   GasW :  18   Gd:241
##  Mean   : 567.2   Mean   :1057.4   Grav :   7   Po:  1
##  3rd Qu.: 808.0   3rd Qu.:1298.2   OthW :   2   TA:428
```

```
##  Max.   :2336.0   Max.   :6110.0   Wall :   4
##
##  Electrical     X1stFlrSF      X2ndFlrSF     LowQualFinSF
##  FuseA:  94   Min.   : 334   Min.   :   0   Min.   :  0.000
##  FuseF:  27   1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000
##  FuseP:   3   Median :1087   Median :   0   Median :  0.000
##  Mix  :   1   Mean   :1163   Mean   : 347   Mean   :  5.845
##  SBrkr:1334   3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000
##  NA's :   1   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   KitchenQual
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex:100
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa: 39
##  Median :0.0000   Median :3.000   Median :1.000   Gd:586
##  Mean   :0.3829   Mean   :2.866   Mean   :1.047   TA:735
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :2.0000   Max.   :8.000   Max.   :3.000
##
##   TotRmsAbvGrd    Functional    Fireplaces    FireplaceQu   GarageType
##  Min.   : 2.000   Maj1:  14   Min.   :0.000   Ex  : 24    2Types :  6
##  1st Qu.: 5.000   Maj2:   5   1st Qu.:0.000   Fa  : 33    Attchd :870
##  Median : 6.000   Min1:  31   Median :1.000   Gd  :380    Basment: 19
##  Mean   : 6.518   Min2:  34   Mean   :0.613   Po  : 20    BuiltIn: 88
##  3rd Qu.: 7.000   Mod :  15   3rd Qu.:1.000   TA  :313    CarPort:  9
##  Max.   :14.000   Sev :   1   Max.   :3.000   NA's:690    Detchd :387
##                   Typ :1360                               NA's   : 81
##   GarageYrBlt   GarageFinish   GarageCars      GarageArea     GarageQual
##  Min.   :1900   Fin :352     Min.   :0.000   Min.   :   0.0   Ex  :   3
##  1st Qu.:1961   RFn :422     1st Qu.:1.000   1st Qu.: 334.5   Fa  :  48
##  Median :1980   Unf :605     Median :2.000   Median : 480.0   Gd  :  14
##  Mean   :1979   NA's: 81     Mean   :1.767   Mean   : 473.0   Po  :   3
##  3rd Qu.:2002                3rd Qu.:2.000   3rd Qu.: 576.0   TA  :1311
##  Max.   :2010                Max.   :4.000   Max.   :1418.0   NA's:  81
##  NA's   :81
##  GarageCond  PavedDrive   WoodDeckSF      OpenPorchSF     EnclosedPorch
##  Ex  :   2   N:  90     Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  Fa  :  35   P:  30     1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
##  Gd  :   9   Y:1340     Median :  0.00   Median : 25.00   Median :  0.00
##  Po  :   7              Mean   : 94.24   Mean   : 46.66   Mean   : 21.95
##  TA  :1326              3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00
##  NA's:  81              Max.   :857.00   Max.   :547.00   Max.   :552.00
##
```

```
##    X3SsnPorch      ScreenPorch        PoolArea        PoolQC
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Ex  :   2
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   Fa  :   2
##  Median :  0.00   Median :  0.00   Median :  0.000   Gd  :   3
##  Mean   :  3.41   Mean   : 15.06   Mean   :  2.759   NA's:1453
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##    Fence      MiscFeature    MiscVal             MoSold
##  GdPrv:  59   Gar2:   2   Min.   :    0.00   Min.   : 1.000
##  GdWo :  54   Othr:   2   1st Qu.:    0.00   1st Qu.: 5.000
##  MnPrv: 157   Shed:  49   Median :    0.00   Median : 6.000
##  MnWw :  11   TenC:   1   Mean   :   43.49   Mean   : 6.322
##  NA's :1179   NA's:1406   3rd Qu.:    0.00   3rd Qu.: 8.000
##                           Max.   :15500.00   Max.   :12.000
##
##      YrSold        SaleType    SaleCondition    SalePrice
##  Min.   :2006   WD     :1267   Abnorml: 101   Min.   : 34900
##  1st Qu.:2007   New    : 122   AdjLand:   4   1st Qu.:129975
##  Median :2008   COD    :  43   Alloca :  12   Median :163000
##  Mean   :2008   ConLD  :   9   Family :  20   Mean   :180921
##  3rd Qu.:2009   ConLI  :   5   Normal :1198   3rd Qu.:214000
##  Max.   :2010   ConLw  :   5   Partial: 125   Max.   :755000
##                 (Other):   9
```

Data Visualization & basic transformation

Before this, going to do some very basic data cleaning

CLEANING DATA

1. Remove ID ID is not really relative data here. It just identifies the rows.

```
housedatac<-select(housedatac, -c(Id))
```

2.. MSSubClass to factor. MSSubclass is actually categorical - represented as numeric. Need to change to factor.

```
housedatac$MSSubClass<-as.factor(housedatac$MSSubClass)
```

MSSubClass: Identifies the type of dwelling involved in the sale.

- 20 1-STORY 1946 & NEWER ALL STYLES
- 30 1-STORY 1945 & OLDER
- 40 1-STORY W/FINISHED ATTIC ALL AGES
- 45 1-1/2 STORY - UNFINISHED ALL AGES
- 50 1-1/2 STORY FINISHED ALL AGES
- 60 2-STORY 1946 & NEWER
- 70 2-STORY 1945 & OLDER
- 75 2-1/2 STORY ALL AGES

- 80 SPLIT OR MULTI-LEVEL
- 85 SPLIT FOYER
- 90 DUPLEX - ALL STYLES AND AGES
- 120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER
- 150 1-1/2 STORY PUD - ALL AGES
- 160 2-STORY PUD - 1946 & NEWER
- 180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
- 190 2 FAMILY CONVERSION - ALL STYLES AND AGES

3. Dividing into numerical and categorical columns to visualize some data.

- Numeric

```
# Numerical into nums
nums <- unlist(lapply(housedatac, is.numeric))
housedata_num<-housedatac[ , nums]

colnames(housedata_num)
```

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

- Categorical

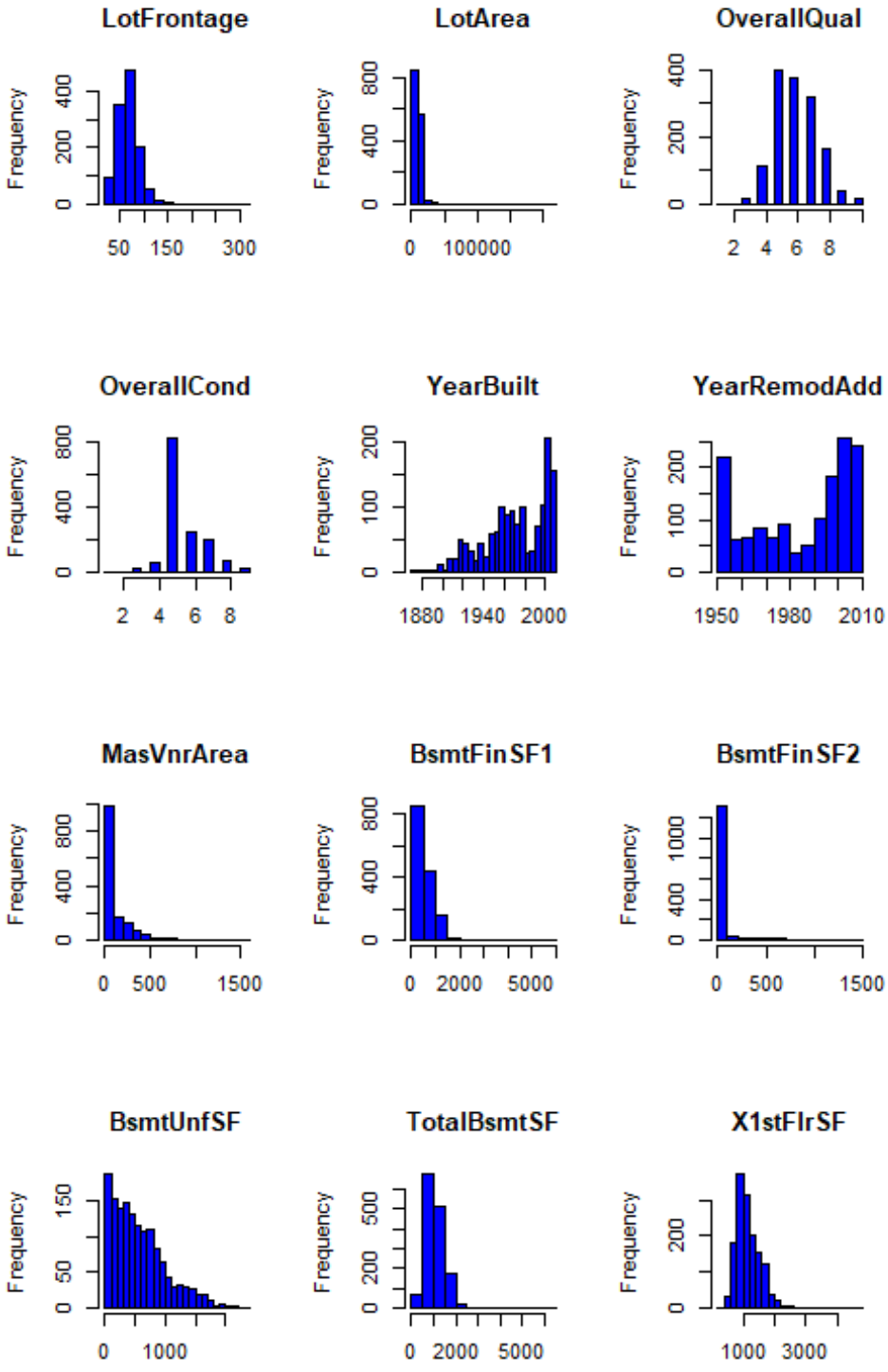
```
# Categorical into cat.

cats <-unlist(lapply(housedatac,is.factor))
housedata_cat<-housedatac[,cats]


colnames(housedata_cat)
```

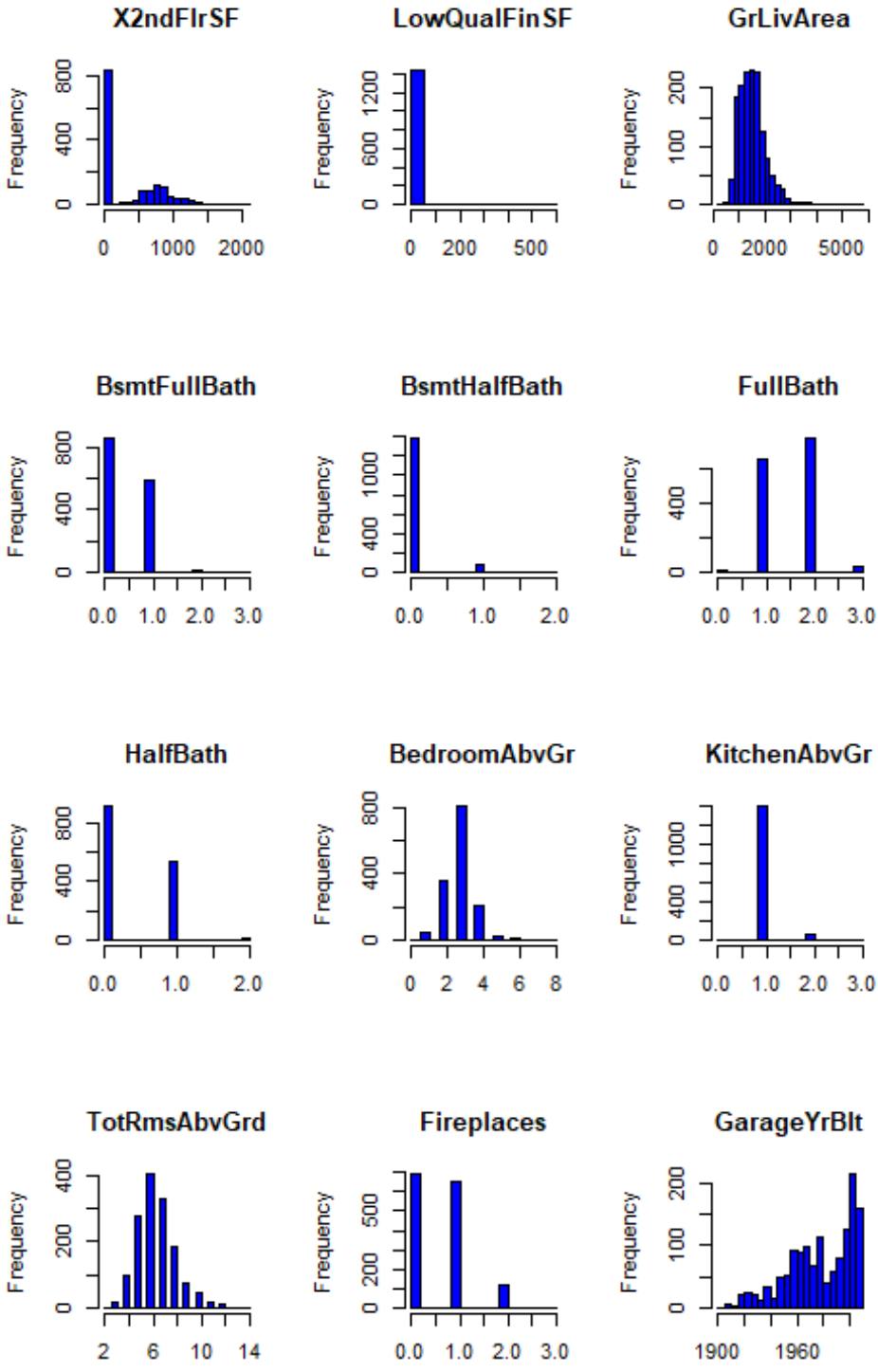
```
##  [1] "MSSubClass"    "MSZoning"      "Street"        "Alley"
##  [5] "LotShape"      "LandContour"   "Utilities"     "LotConfig"
##  [9] "LandSlope"     "Neighborhood"  "Condition1"    "Condition2"
## [13] "BldgType"      "HouseStyle"    "RoofStyle"     "RoofMatl"
## [17] "Exterior1st"   "Exterior2nd"   "MasVnrType"    "ExterQual"
## [21] "ExterCond"     "Foundation"    "BsmtQual"      "BsmtCond"
## [25] "BsmtExposure"  "BsmtFinType1"  "BsmtFinType2"  "Heating"
## [29] "HeatingQC"     "CentralAir"    "Electrical"    "KitchenQual"
## [33] "Functional"    "FireplaceQu"   "GarageType"    "GarageFinish"
## [37] "GarageQual"    "GarageCond"    "PavedDrive"    "PoolQC"
## [41] "Fence"         "MiscFeature"   "SaleType"      "SaleCondition"
```

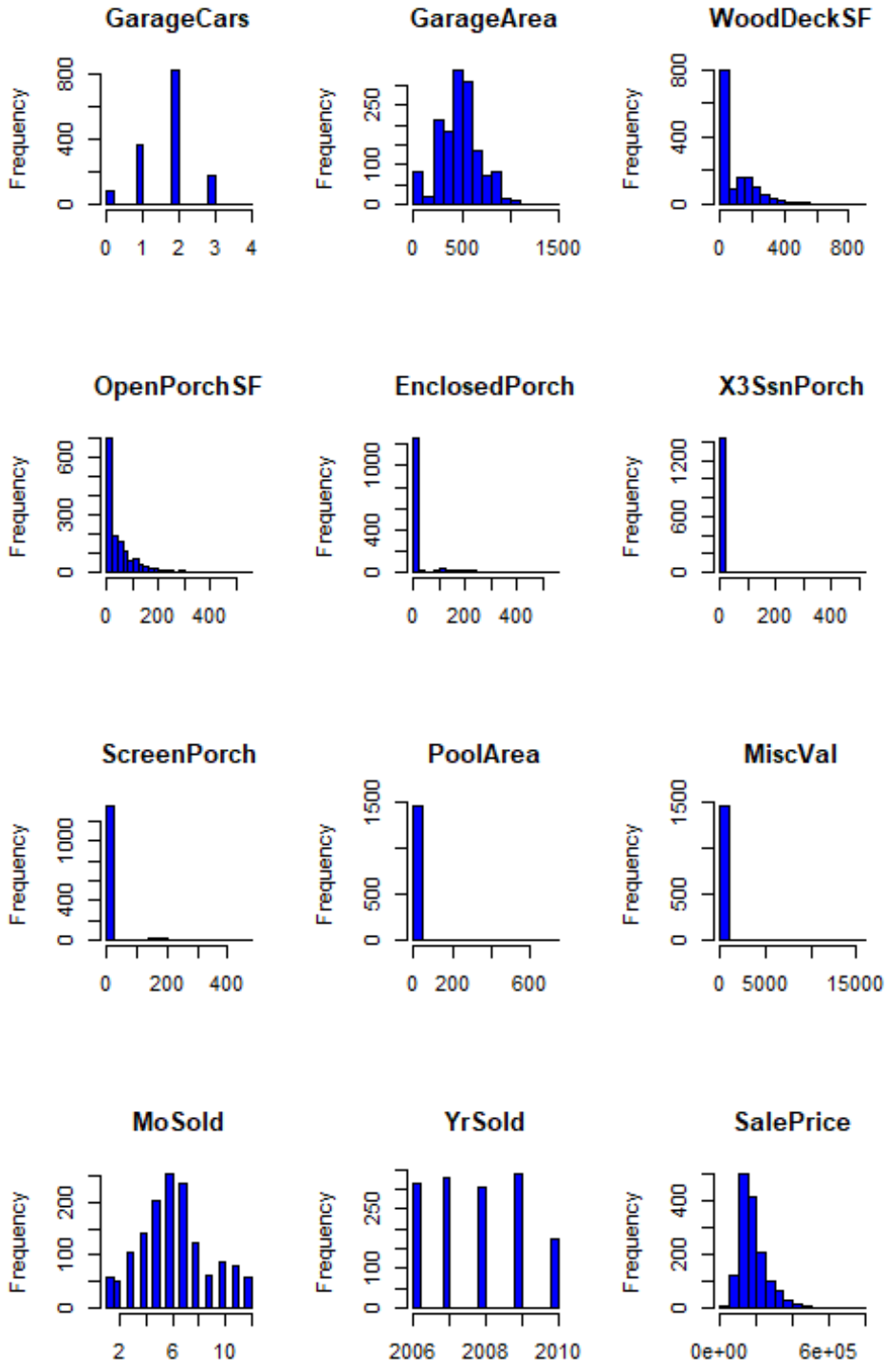
In step2, will go through ordinal variables and make some changes to this structure.

Plot numerical - Histogram

```
#sapply(housedata_num, hist,main=colnames(housedata_num), )
par(mfrow=c(2,3))
mapply(hist,as.data.frame(housedata_num),main=colnames(housedata_num),xlab=''
,breaks=20, col="blue")
```


LotFrontage
Frequency
0 200 400
50 150 300
LotArea
Frequency
0 400 800
0 100000
OverallQual
Frequency
0 200 400
2 4 6 8
OverallCond
Frequency
0 400 800
2 4 6 8
YearBuilt
Frequency
0 100 200
1880 1940 2000
YearRemodAdd
Frequency
0 100 200
1950 1980 2010
MasVnrArea
Frequency
0 400 800
0 500 1500
BsmtFinSF1
Frequency
0 400 800
0 2000 5000
BsmtFinSF2
Frequency
0 400 1000
0 500 1500
BsmtUnfSF
Frequency
0 50 150
0 1000
TotalBsmtSF
Frequency
0 200 500
0 2000 5000
X1stFlrSF
Frequency
0 200
1000 3000


X2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageYrBlt
Frequency




```
##           LotFrontage        LotArea            OverallQual
## breaks    Integer,16         Numeric,23         Numeric,19
## counts    Integer,15         Integer,22         Integer,18
## density   Numeric,15         Numeric,22         Numeric,18
```

```
## mids     Numeric,15          Numeric,22          Numeric,18
## xname    "dots[[1L]][[1L]]" "dots[[1L]][[2L]]" "dots[[1L]][[3L]]"
## equidist TRUE                TRUE                TRUE
##          OverallCond         YearBuilt           YearRemodAdd
## breaks   Numeric,17          Integer,29          Integer,13
## counts   Integer,16          Integer,28          Integer,12
## density  Numeric,16          Numeric,28          Numeric,12
## mids     Numeric,16          Numeric,28          Numeric,12
## xname    "dots[[1L]][[4L]]" "dots[[1L]][[5L]]" "dots[[1L]][[6L]]"
## equidist TRUE                TRUE                TRUE
##          MasVnrArea          BsmtFinSF1          BsmtFinSF2
## breaks   Numeric,17          Numeric,13          Numeric,16
## counts   Integer,16          Integer,12          Integer,15
## density  Numeric,16          Numeric,12          Numeric,15
## mids     Numeric,16          Numeric,12          Numeric,15
## xname    "dots[[1L]][[7L]]" "dots[[1L]][[8L]]" "dots[[1L]][[9L]]"
## equidist TRUE                TRUE                TRUE
##          BsmtUnfSF            TotalBsmtSF           X1stFlrSF
## breaks   Numeric,25           Numeric,14            Integer,24
## counts   Integer,24           Integer,13            Integer,23
## density  Numeric,24           Numeric,13            Numeric,23
## mids     Numeric,24           Numeric,13            Numeric,23
## xname    "dots[[1L]][[10L]]" "dots[[1L]][[11L]]" "dots[[1L]][[12L]]"
## equidist TRUE                TRUE                TRUE
##          X2ndFlrSF           LowQualFinSF        GrLivArea
## breaks   Numeric,22          Numeric,13          Integer,29
## counts   Integer,21          Integer,12          Integer,28
## density  Numeric,21          Numeric,12          Numeric,28
## mids     Numeric,21          Numeric,12          Numeric,28
## xname    "dots[[1L]][[13L]]" "dots[[1L]][[14L]]" "dots[[1L]][[15L]]"
## equidist TRUE                TRUE                TRUE
##          BsmtFullBath        BsmtHalfBath        FullBath
## breaks   Numeric,16          Numeric,21          Numeric,16
## counts   Integer,15          Integer,20          Integer,15
## density  Numeric,15          Numeric,20          Numeric,15
## mids     Numeric,15          Numeric,20          Numeric,15
## xname    "dots[[1L]][[16L]]" "dots[[1L]][[17L]]" "dots[[1L]][[18L]]"
## equidist TRUE                TRUE                TRUE
##          HalfBath            BedroomAbvGr        KitchenAbvGr
## breaks   Numeric,21          Numeric,17          Numeric,16
## counts   Integer,20          Integer,16          Integer,15
## density  Numeric,20          Numeric,16          Numeric,15
## mids     Numeric,20          Numeric,16          Numeric,15
## xname    "dots[[1L]][[19L]]" "dots[[1L]][[20L]]" "dots[[1L]][[21L]]"
## equidist TRUE                TRUE                TRUE
##          TotRmsAbvGrd        Fireplaces          GarageYrBlt
## breaks   Numeric,25          Numeric,16          Integer,23
## counts   Integer,24          Integer,15          Integer,22
## density  Numeric,24          Numeric,15          Numeric,22
## mids     Numeric,24          Numeric,15          Numeric,22
```

```
## xname    "dots[[1L]][[22L]]" "dots[[1L]][[23L]]" "dots[[1L]][[24L]]"
## equidist TRUE                TRUE                TRUE
##          GarageCars          GarageArea          WoodDeckSF
## breaks   Numeric,21          Numeric,16          Numeric,19
## counts   Integer,20          Integer,15          Integer,18
## density  Numeric,20          Numeric,15          Numeric,18
## mids     Numeric,20          Numeric,15          Numeric,18
## xname    "dots[[1L]][[25L]]" "dots[[1L]][[26L]]" "dots[[1L]][[27L]]"
## equidist TRUE                TRUE                TRUE
##          OpenPorchSF         EnclosedPorch       X3SsnPorch
## breaks   Numeric,29          Numeric,29          Numeric,27
## counts   Integer,28          Integer,28          Integer,26
## density  Numeric,28          Numeric,28          Numeric,26
## mids     Numeric,28          Numeric,28          Numeric,26
## xname    "dots[[1L]][[28L]]" "dots[[1L]][[29L]]" "dots[[1L]][[30L]]"
## equidist TRUE                TRUE                TRUE
##          ScreenPorch         PoolArea            MiscVal
## breaks   Numeric,25          Numeric,16          Numeric,17
## counts   Integer,24          Integer,15          Integer,16
## density  Numeric,24          Numeric,15          Numeric,16
## mids     Numeric,24          Numeric,15          Numeric,16
## xname    "dots[[1L]][[31L]]" "dots[[1L]][[32L]]" "dots[[1L]][[33L]]"
## equidist TRUE                TRUE                TRUE
##          MoSold              YrSold              SalePrice
## breaks   Numeric,23          Numeric,21          Numeric,17
## counts   Integer,22          Integer,20          Integer,16
## density  Numeric,22          Numeric,20          Numeric,16
## mids     Numeric,22          Numeric,20          Numeric,16
## xname    "dots[[1L]][[34L]]" "dots[[1L]][[35L]]" "dots[[1L]][[36L]]"
## equidist TRUE                TRUE                TRUE
```

Many of the distributions are skewed - including salePrice Log transformation would be beneficial.

Skewness Test

Originally used Shapiro wilk test, but this only works well with n<50 and thus all my distributions showed as not-normal. Skewness - between -0.5 to 0.5 normal, absolute value between 0.5 and 1 moderately skewed. Abs value above 1 is highly skewed. Kurtosis - Kurtosis tells you the height and sharpness of the central peak, relative to that of a standard bell curve.

```
#skewness.test(housedata_num$SalePrice)
Check<-describe(housedata_num)[,c(11,12)]
Check
```

```
##                skew kurtosis
## LotFrontage    2.16    17.34
## LotArea       12.18   202.26
## OverallQual    0.22     0.09
```

```
## OverallCond    0.69     1.09
## YearBuilt     -0.61    -0.45
## YearRemodAdd  -0.50    -1.27
## MasVnrArea     2.66    10.03
## BsmtFinSF1     1.68    11.06
## BsmtFinSF2     4.25    20.01
## BsmtUnfSF      0.92     0.46
## TotalBsmtSF    1.52    13.18
## X1stFlrSF      1.37     5.71
## X2ndFlrSF      0.81    -0.56
## LowQualFinSF   8.99    82.83
## GrLivArea      1.36     4.86
## BsmtFullBath   0.59    -0.84
## BsmtHalfBath   4.09    16.31
## FullBath       0.04    -0.86
## HalfBath       0.67    -1.08
## BedroomAbvGr   0.21     2.21
## KitchenAbvGr   4.48    21.42
## TotRmsAbvGrd   0.67     0.87
## Fireplaces     0.65    -0.22
## GarageYrBlt   -0.65    -0.42
## GarageCars    -0.34     0.21
## GarageArea     0.18     0.90
## WoodDeckSF     1.54     2.97
## OpenPorchSF    2.36     8.44
## EnclosedPorch  3.08    10.37
## X3SsnPorch    10.28   123.06
## ScreenPorch    4.11    18.34
## PoolArea      14.80   222.19
## MiscVal       24.43   697.64
## MoSold         0.21    -0.41
## YrSold         0.10    -1.19
## SalePrice      1.88     6.50
```

Below items all high skewness and may need transformation.

- LotFrontage 2.16 17.34
- LotArea 12.18 202.26
- MasVnrArea 2.66 10.03
- BsmtFinSF1 1.68 11.06
- BsmtFinSF2 4.25 20.01
- TotalBsmtSF 1.52 13.18
- X1stFlrSF 1.37 5.71
- LowQualFinSF 8.99 82.83
- GrLivArea 1.36 4.86
- BsmtHalfBath 4.09 16.31
- KitchenAbvGr 4.48 21.42
- OpenPorchSF 2.36 8.44

- EnclosedPorch 3.08 10.37
- X3SsnPorch 10.28 123.06
- ScreenPorch 4.11 18.34
- PoolArea 14.80 222.19
- MiscVal 24.43 697.64
- SalePrice 1.88 6.50

Subset non-normal variables.

```
H_SkewedVar<-select(housedata_num,c(


LotArea,
MasVnrArea,
BsmtFinSF1,
BsmtFinSF2,
TotalBsmtSF,
X1stFlrSF,
LowQualFinSF,
GrLivArea,
BsmtHalfBath,
KitchenAbvGr,
OpenPorchSF,
EnclosedPorch,
X3SsnPorch,
ScreenPorch,
PoolArea,
MiscVal,
SalePrice))
```

Apply Log base 10 and check skewness.

(log10(x+1) because there are some x=0 values)

```
describe(H_SkewedVar)[,c(11,12)]
```

```
##                skew kurtosis
## LotArea       12.18   202.26
## MasVnrArea     2.66    10.03
## BsmtFinSF1     1.68    11.06
## BsmtFinSF2     4.25    20.01
## TotalBsmtSF    1.52    13.18
## X1stFlrSF      1.37     5.71
## LowQualFinSF   8.99    82.83
## GrLivArea      1.36     4.86
## BsmtHalfBath   4.09    16.31
## KitchenAbvGr   4.48    21.42
## OpenPorchSF    2.36     8.44
## EnclosedPorch  3.08    10.37
## X3SsnPorch    10.28   123.06
```

```
## ScreenPorch    4.11    18.34
## PoolArea      14.80   222.19
## MiscVal       24.43   697.64
## SalePrice      1.88     6.50
```

```
describe(log10(H_SkewedVar+1),)[,c(11,12)]
```

```
##                skew kurtosis
## LotArea       -0.14     4.68
## MasVnrArea     0.49    -1.63
## BsmtFinSF1    -0.62    -1.48
## BsmtFinSF2     2.52     4.52
## TotalBsmtSF   -5.14    27.60
## X1stFlrSF      0.08     0.14
## LowQualFinSF   7.44    54.12
## GrLivArea     -0.01     0.27
## BsmtHalfBath   3.92    13.78
## KitchenAbvGr   3.86    19.95
## OpenPorchSF   -0.02    -1.78
## EnclosedPorch  2.11     2.55
## X3SsnPorch     7.72    58.13
## ScreenPorch    3.14     7.98
## PoolArea      14.33   203.69
## MiscVal        5.16    25.31
## SalePrice      0.12     0.80
```

The Log transformation significantly improved skewness for below.

```
            skew kurtosis
```

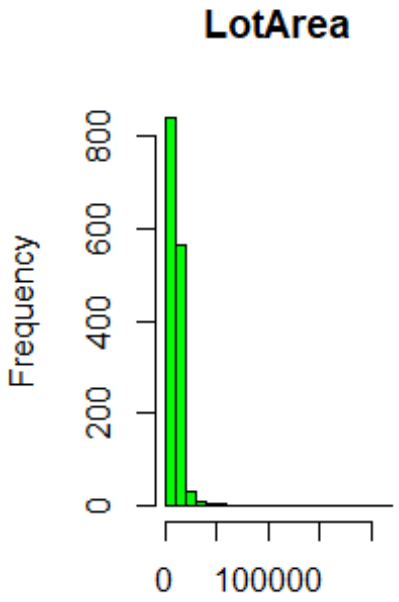
- LotArea -0.14 4.68
- MasVnrArea 0.49 -1.63
- BsmtFinSF1 -0.62 -1.48
- BsmtFinSF2 2.52 4.52
- X1stFlrSF 0.08 0.14
- GrLivArea -0.01 0.27
- OpenPorchSF -0.02 -1.78
- SalePrice 0.12 0.80

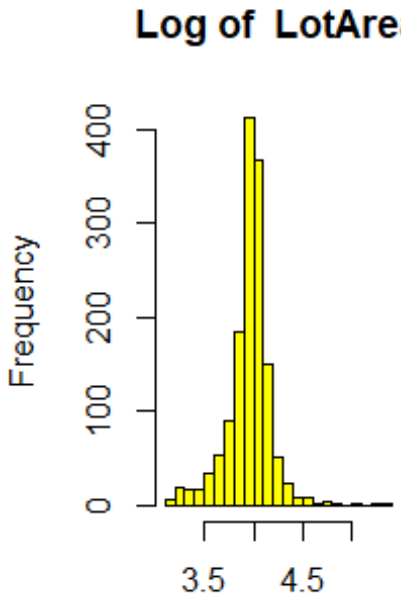
Below didn't help much with log transformation

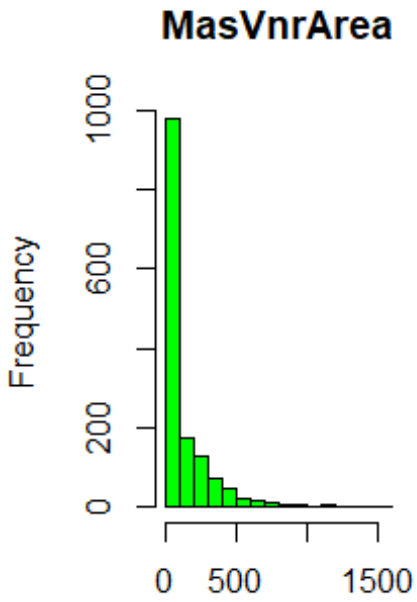
- LowQualFinSF 7.44
- BsmtHalfBath 3.92 13.78
- KitchenAbvGr 4.48 21.42
- X3SsnPorch 7.72 58.13
- ScreenPorch 3.14 7.98
- PoolArea 14.33 203.69
- MiscVal 5.16 25.31

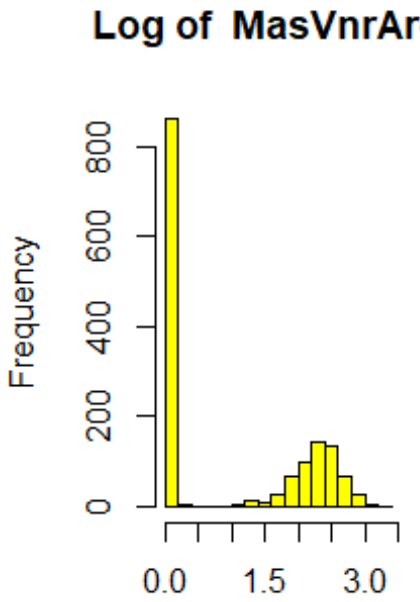
Plot Histogram to see if this matches with the skewness changes.

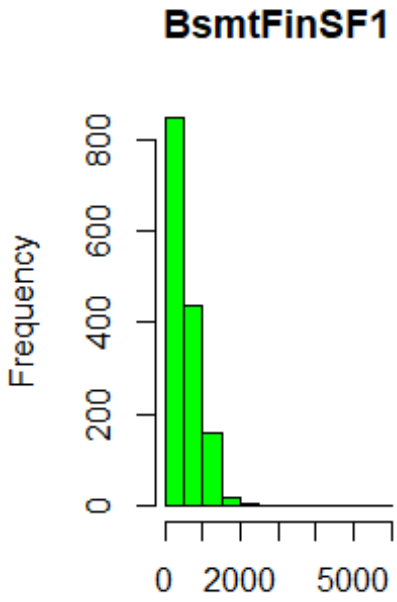
```
par(mfrow=c(1,2))
for (i in 1:17){
hist(H_SkewedVar[,i],main=colnames(H_SkewedVar[i]),xlab='' ,breaks=20,
col="green")
#  lines(density(H_SkewedVar[,i], na.rm=T),col="green")
hist(log10(H_SkewedVar[,i]+1),main=paste("Log of
",colnames(H_SkewedVar[i])),xlab='' ,breaks=20, col="yellow")
# lines(density(log10(H_SkewedVar[,i]), na.rm=T),col="yellow")
}
```

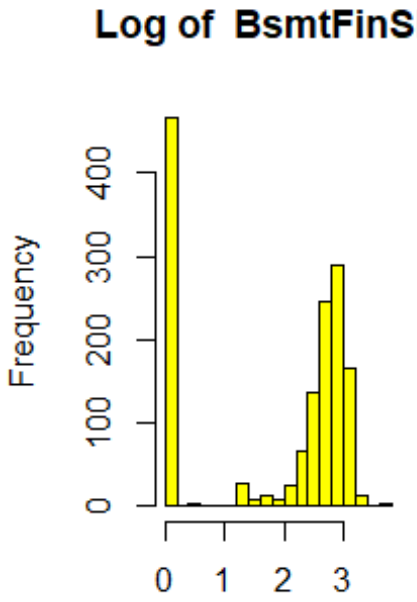

LotArea
Frequency
0
200
400
600
800
0
100000

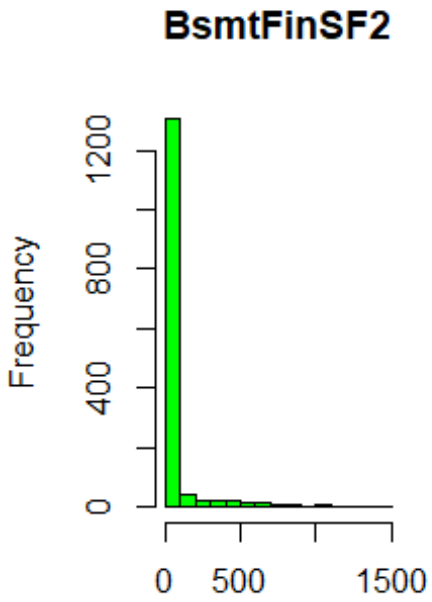

Log of LotArea
Frequency
0
100
200
300
400
3.5
4.5

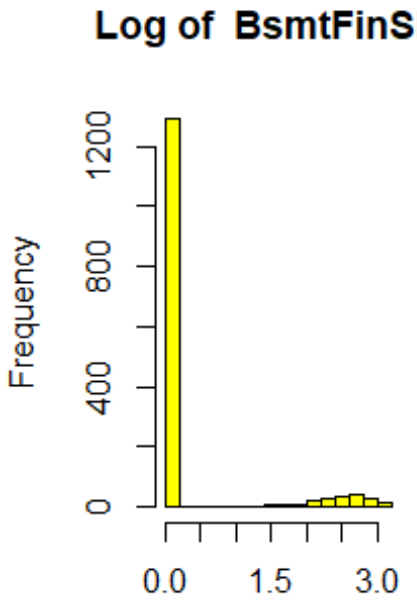

MasVnrArea
Frequency
0
200
600
1000
0
500
1500

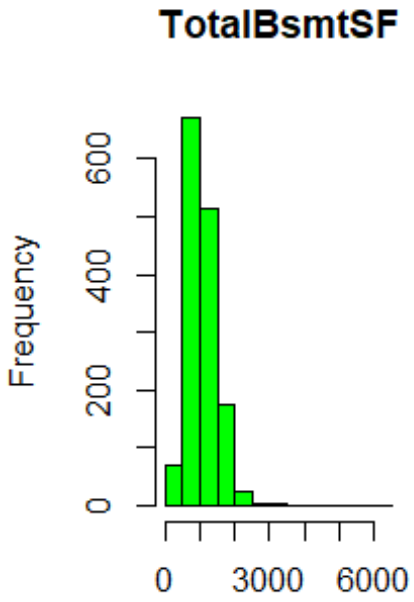

Log of MasVnrArea
Frequency
0
200
400
600
800
0.0
1.5
3.0

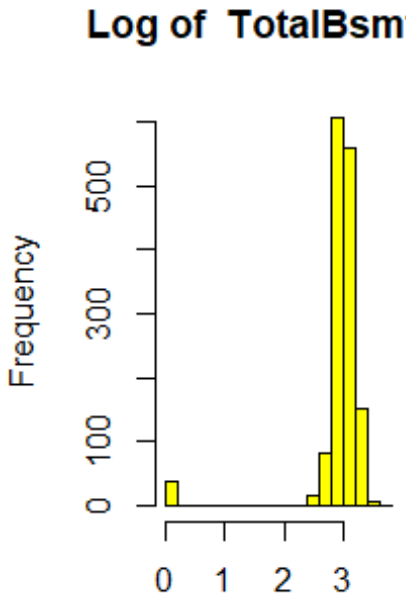

BsmtFinSF1
Frequency
0
200
400
600
800
0
2000
5000

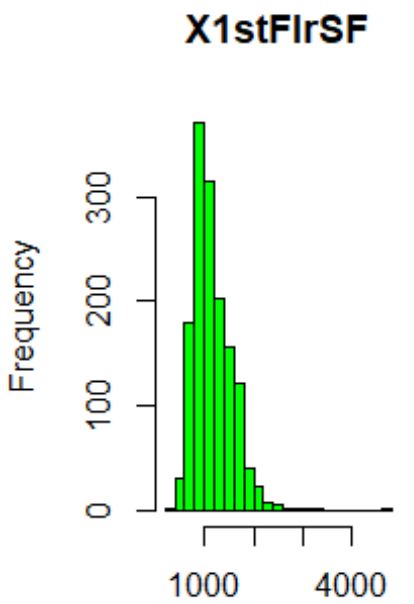

Log of BsmtFinSF1
Frequency
0
100
200
300
400
0
1
2
3

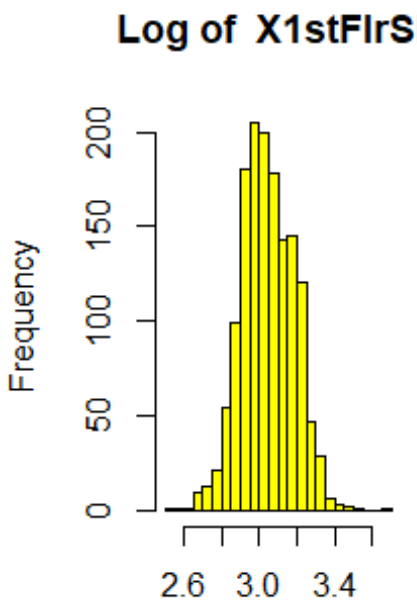

BsmtFinSF2
Frequency
0
400
800
1200
0
500
1500

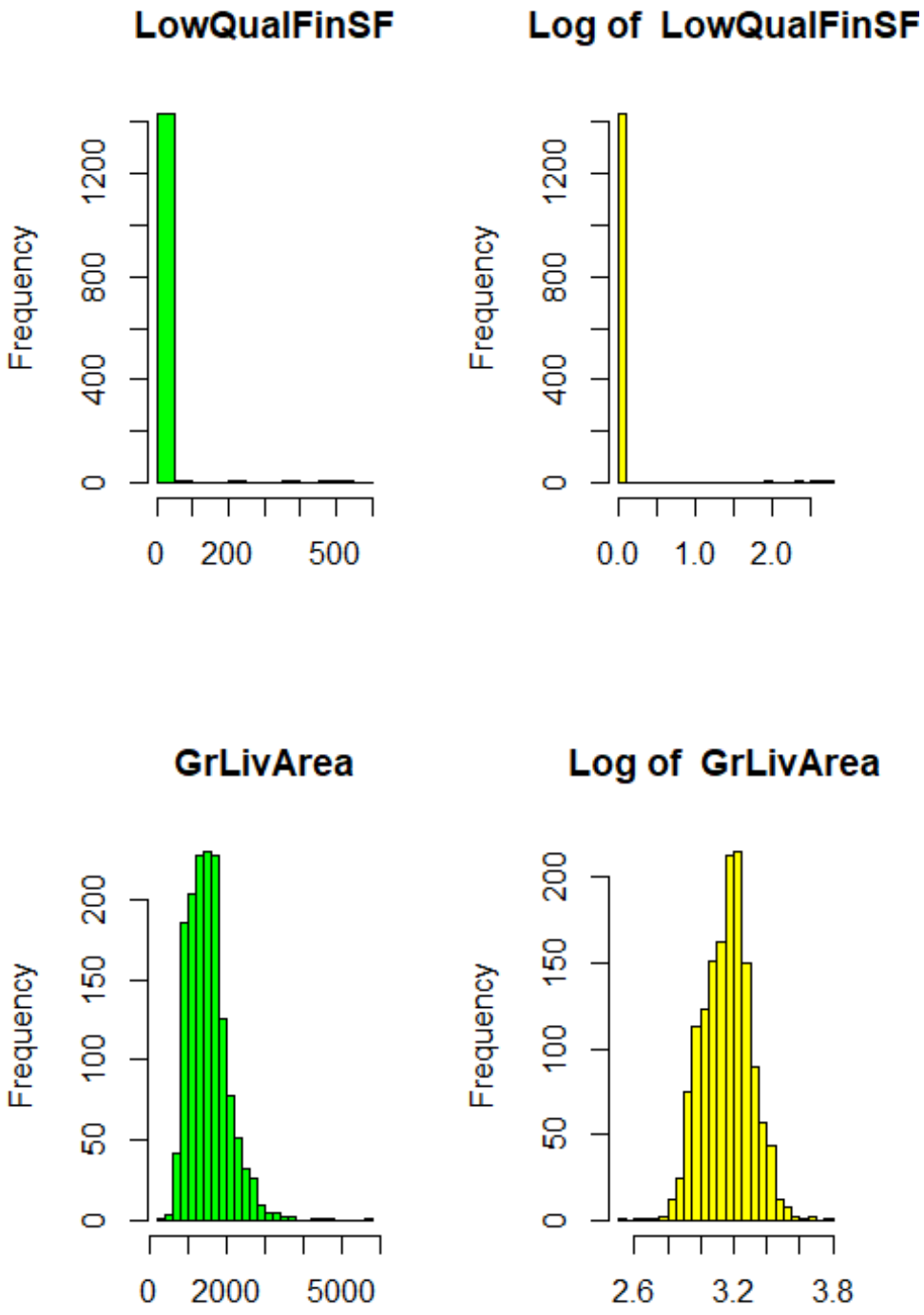

Log of BsmtFinSF2
Frequency
0
400
800
1200
0.0
1.5
3.0

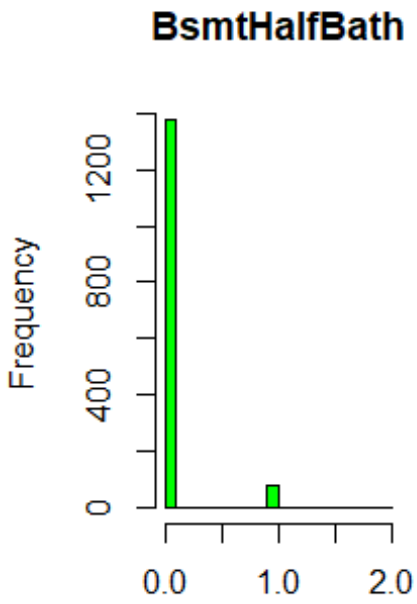

TotalBsmtSF
Frequency
0
200
400
600
0
3000
6000

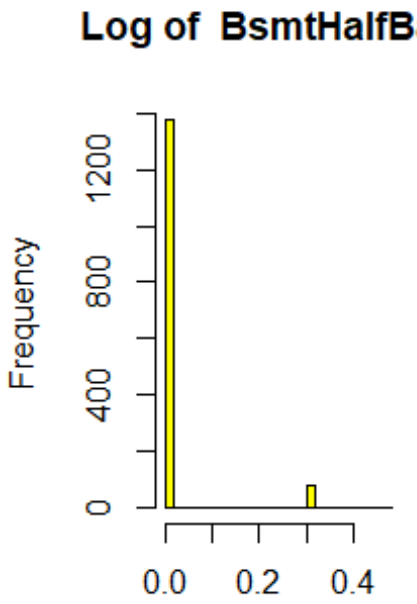

Log of TotalBsmtSF
Frequency
0
100
300
500
0
1
2
3

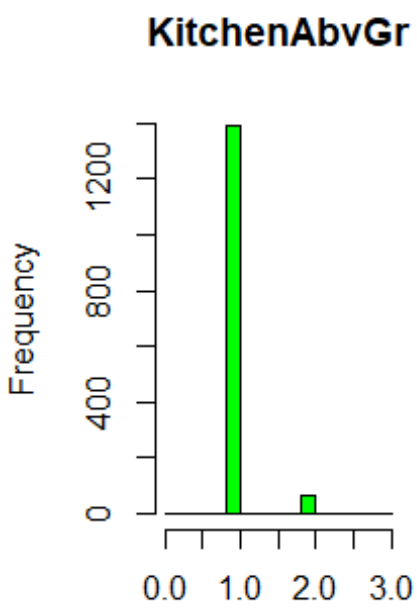

X1stFlrSF
Frequency
0
100
200
300
1000
4000

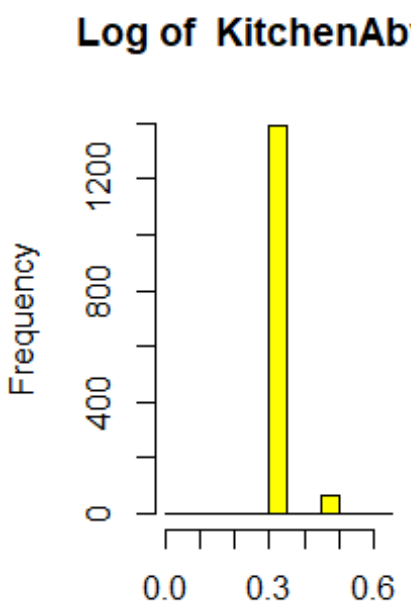

Log of X1stFlrSF
Frequency
0
50
100
150
200
2.6
3.0
3.4

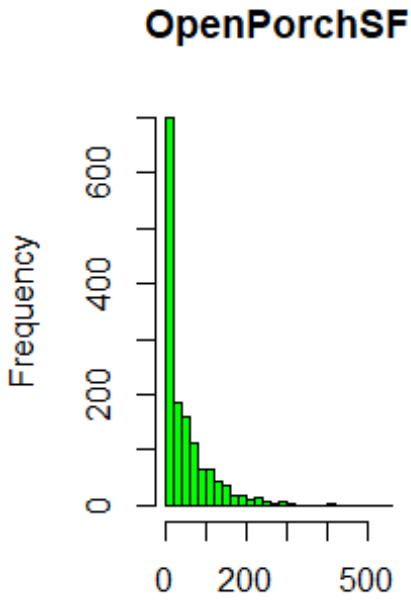

LowQualFinSF
Frequency
0
400
800
1200
0
200
500
Log of LowQualFinSF
Frequency
0
400
800
1200
0.0
1.0
2.0
GrLivArea
Frequency
0
50
100
150
200
0
2000
5000
Log of GrLivArea
Frequency
0
50
100
150
200
2.6
3.2
3.8

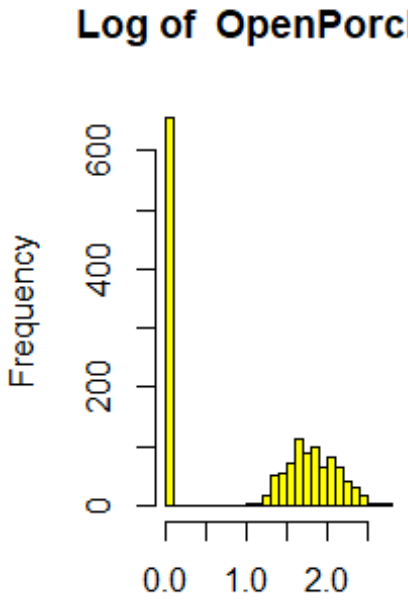

BsmtHalfBath
Frequency
0
400
800
1200
0.0
1.0
2.0

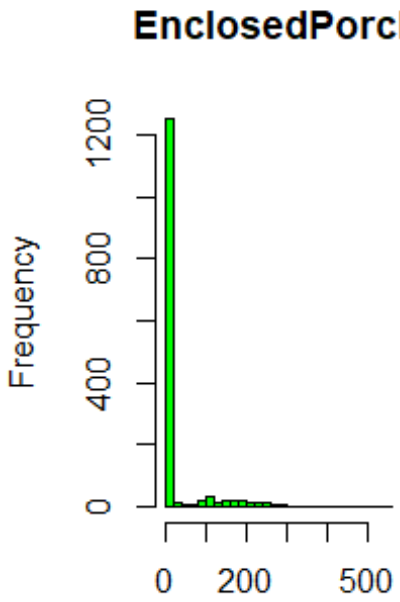

Log of BsmtHalfBath
Frequency
0
400
800
1200
0.0
0.2
0.4

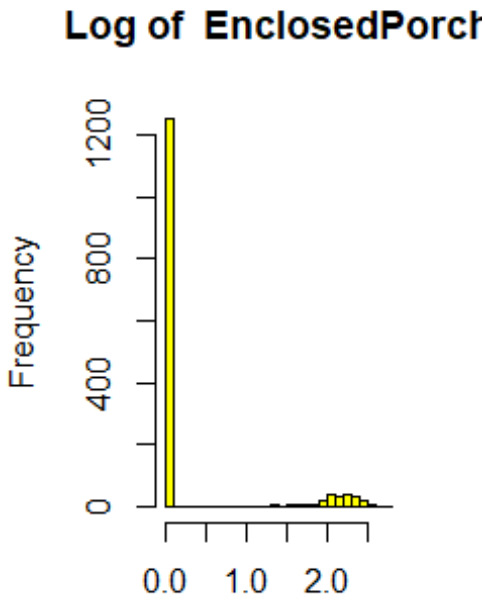

KitchenAbvGr
Frequency
0
400
800
1200
0.0
1.0
2.0
3.0

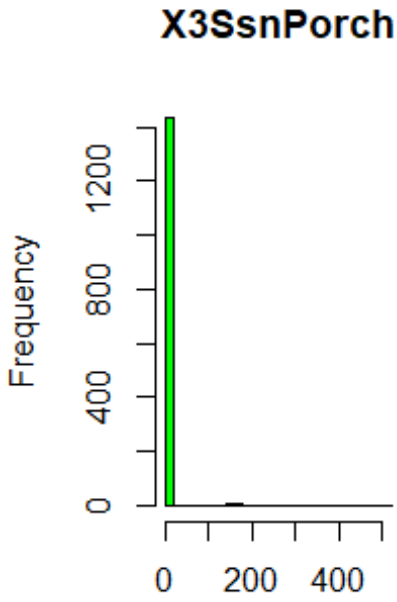

Log of KitchenAbvGr
Frequency
0
400
800
1200
0.0
0.3
0.6

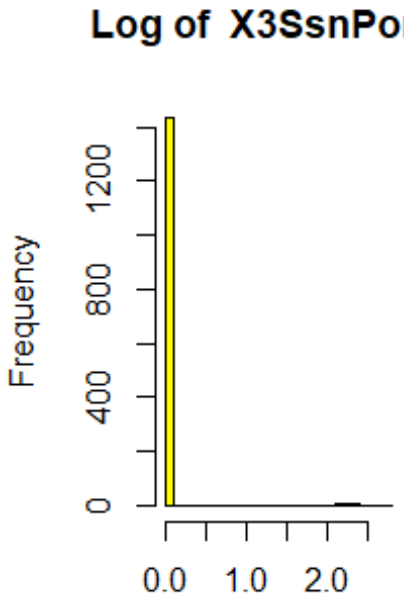

OpenPorchSF
Frequency
0
200
400
600
0
200
500

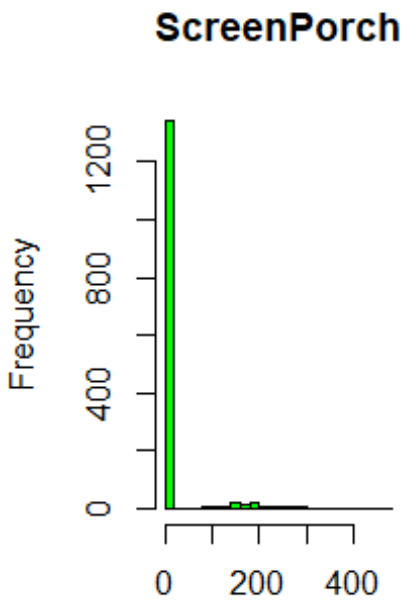

Log of OpenPorchSF
Frequency
0
200
400
600
0.0
1.0
2.0

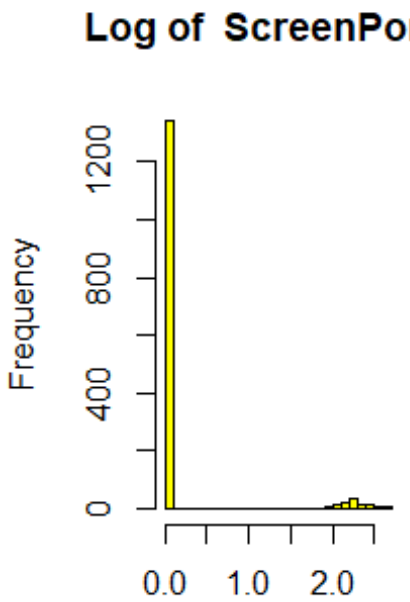

EnclosedPorch
Frequency
0
400
800
1200
0
200
500

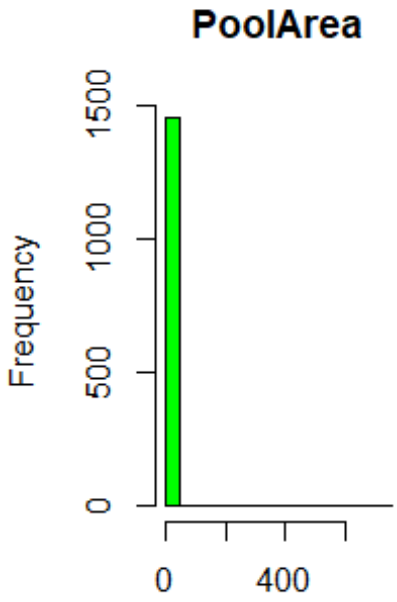

Log of EnclosedPorch
Frequency
0
400
800
1200
0.0
1.0
2.0

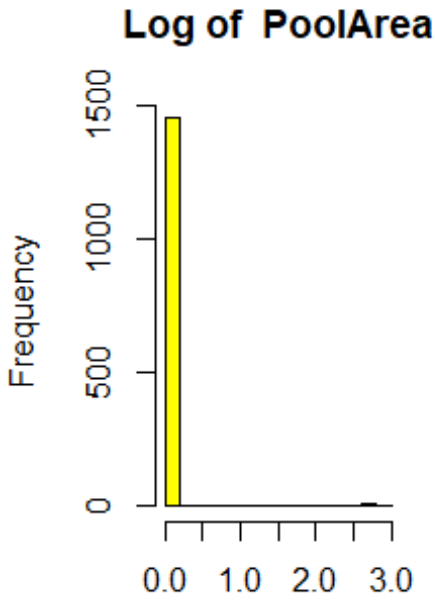

X3SsnPorch
Frequency
0
400
800
1200
0
200
400

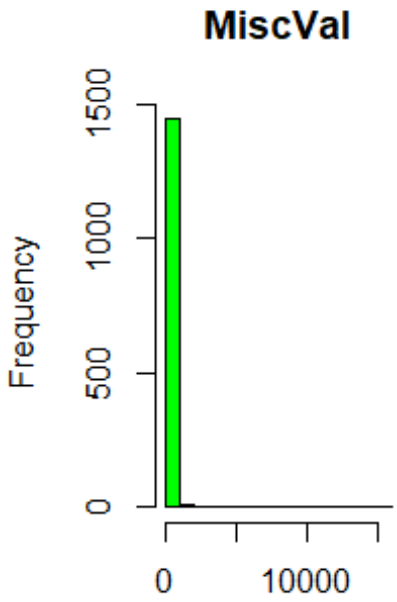

Log of X3SsnPorch
Frequency
0
400
800
1200
0.0
1.0
2.0

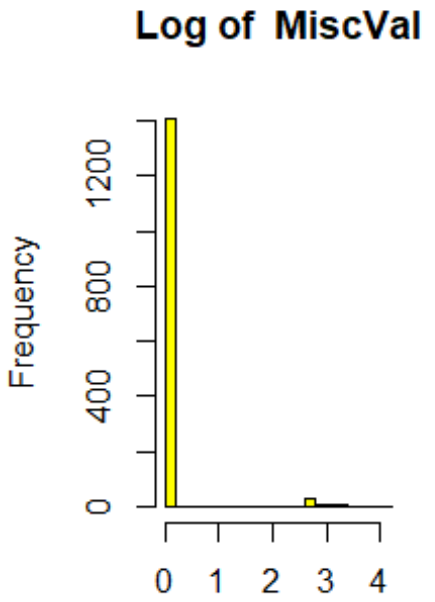

ScreenPorch
Frequency
0
400
800
1200
0
200
400

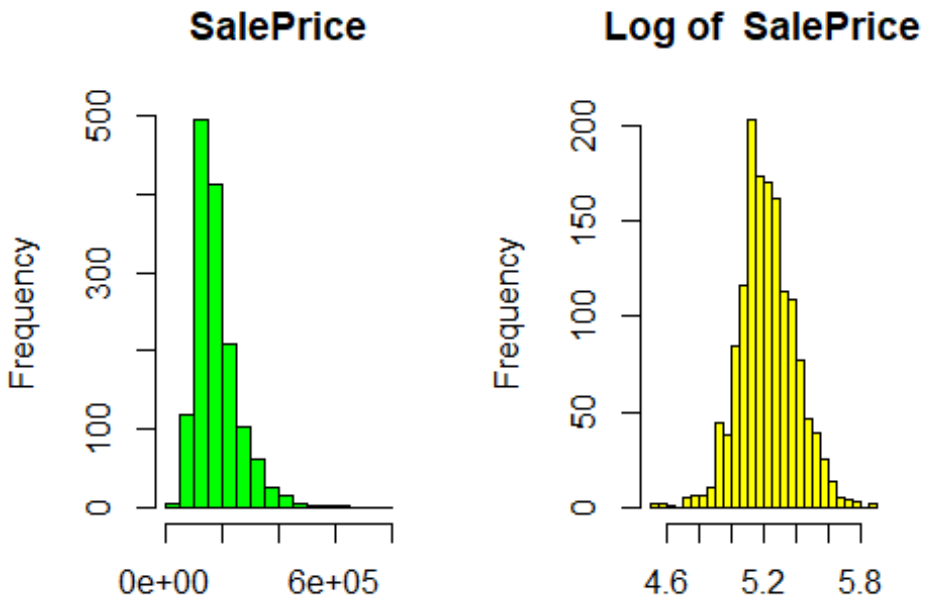

Log of ScreenPorch
Frequency
0
400
800
1200
0.0
1.0
2.0


PoolArea
Frequency
0
500
1000
1500
0
400


Log of PoolArea
Frequency
0
500
1000
1500
0.0
1.0
2.0
3.0


MiscVal
Frequency
0
500
1000
1500
0
10000


Log of MiscVal
Frequency
0
400
800
1200
0
1
2
3
4



Looks like it matches with the skewness values.

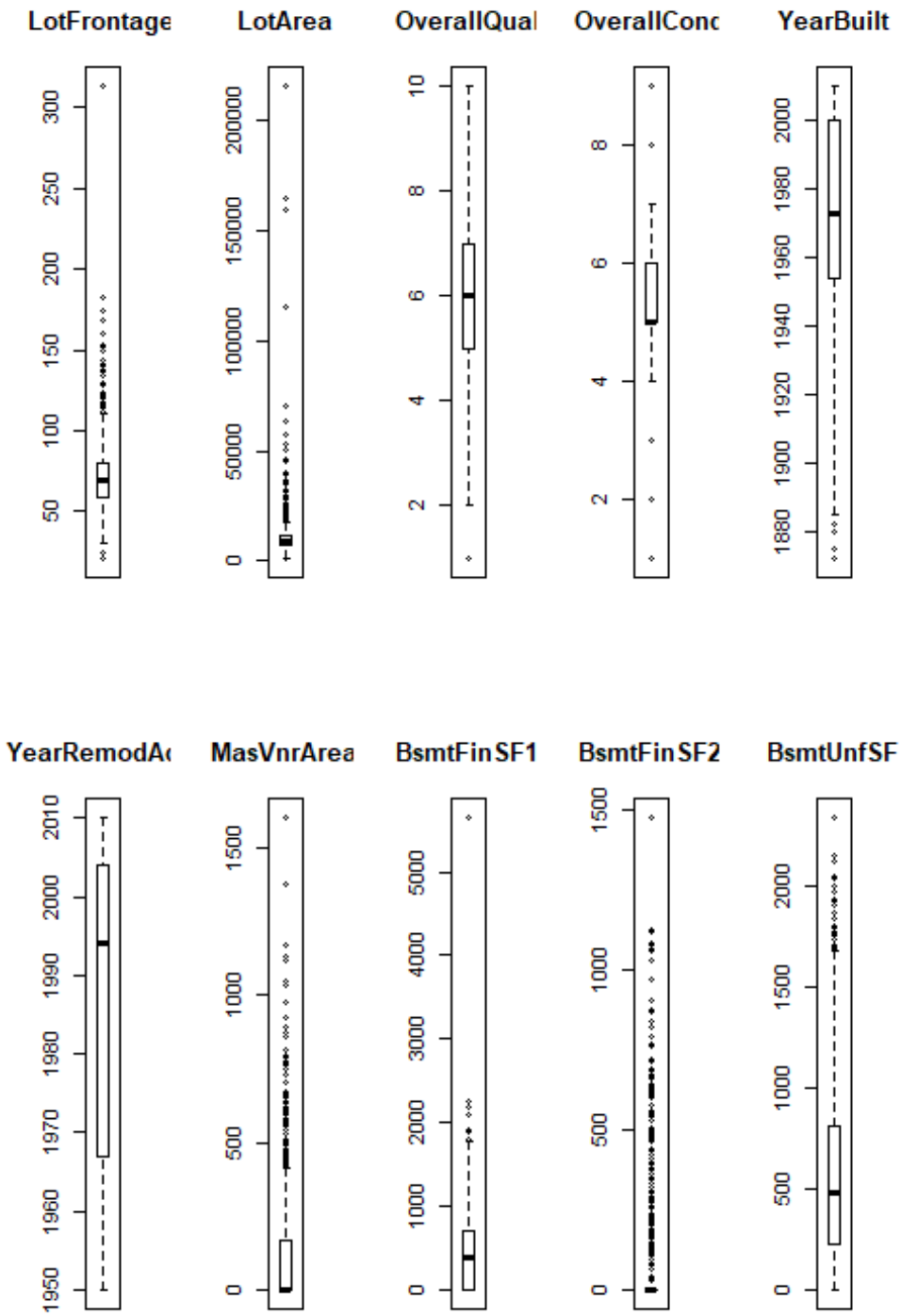
```
           skew kurtosis
```

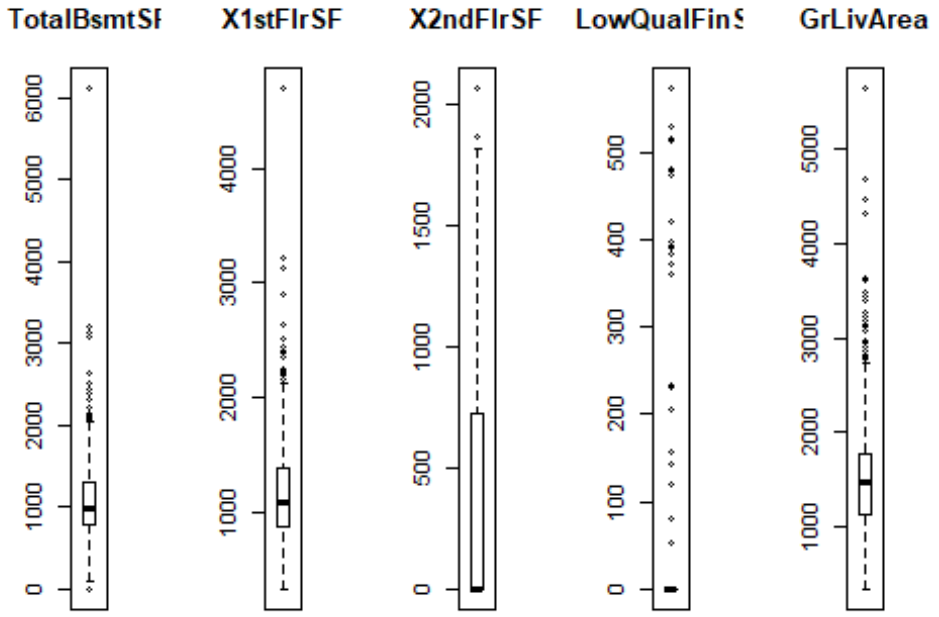
LotArea -0.14 4.68 MasVnrArea 0.49 -1.63 BsmtFinSF1 -0.62 -1.48 BsmtFinSF2 2.52 4.52 X1stFlrSF 0.08 0.14 GrLivArea -0.01 0.27 OpenPorchSF -0.02 -1.78 SalePrice 0.12 0.80

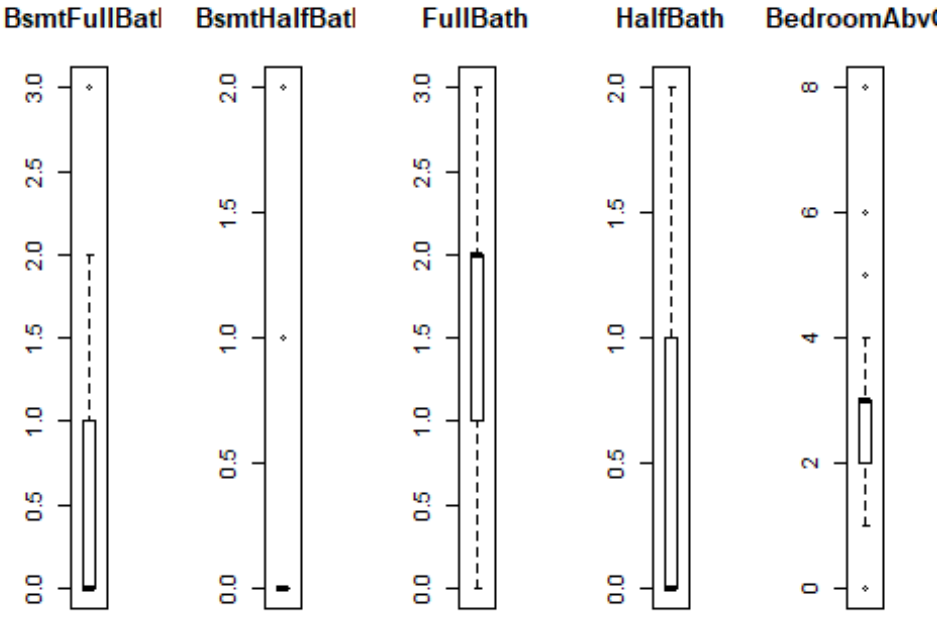
Plot numerical 2 - Boxplot to check outliers

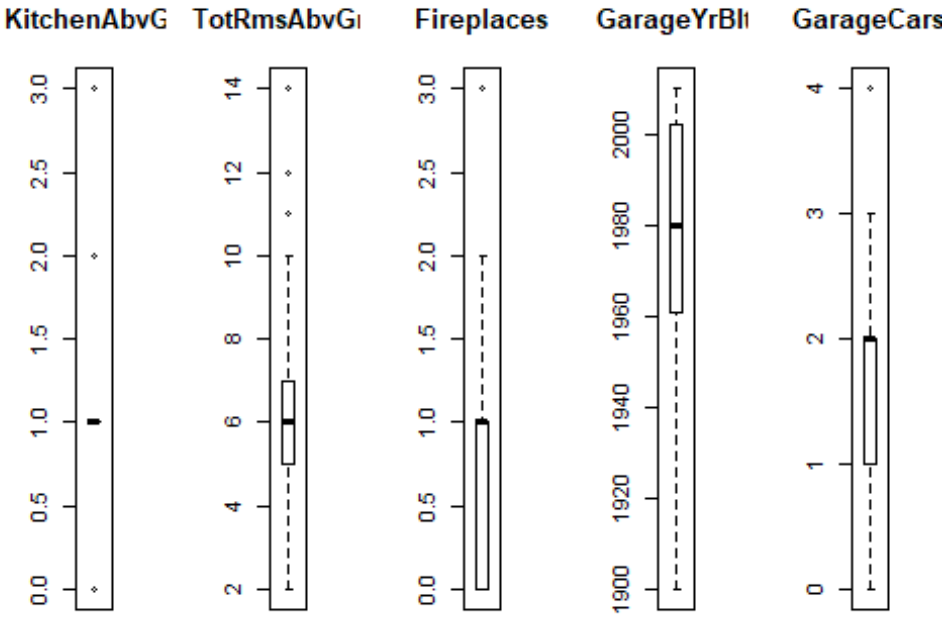
An outlier is defined as above 1.5 interqurtile range above the upper quartile(Q3) and below 1.5 of lower quartile(Q1). Looking at the boxplot, there are too many outliers. Many are because the distribution is skewed, and some because majority of the distribution are packed into very small range (variance is very small).

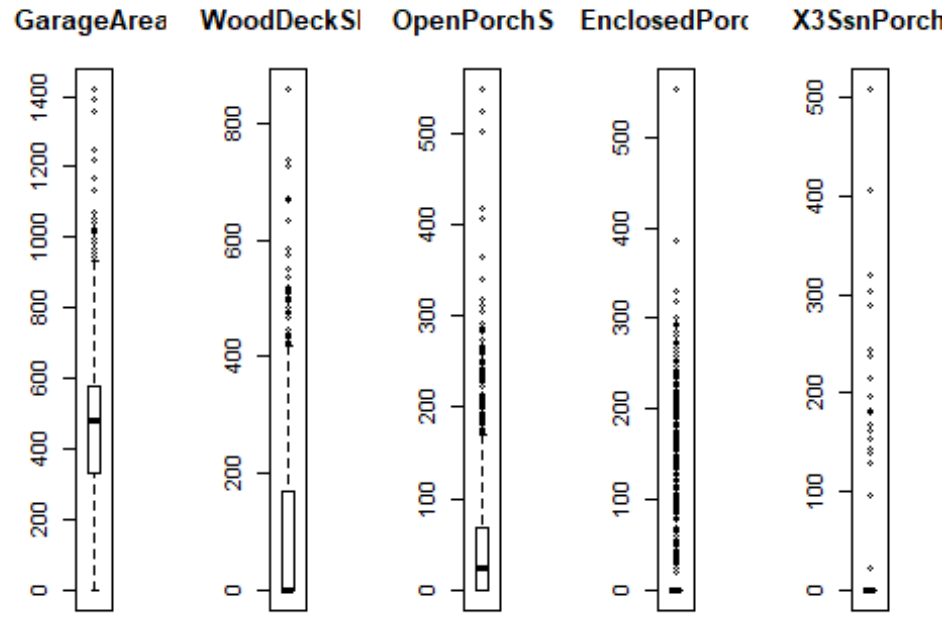
```
par(mfrow=c(1,5))


# The mapply() function stands for 'multivariate' apply. Its purpose is to be
able to vectorize arguments to a function that is not usually accepting
vectors as arguments. In short, mapply() applies a Function to Multiple List
or multiple Vector Arguments.

mapply(boxplot,data.frame(housedata_num),main=colnames(housedata_num) )
```

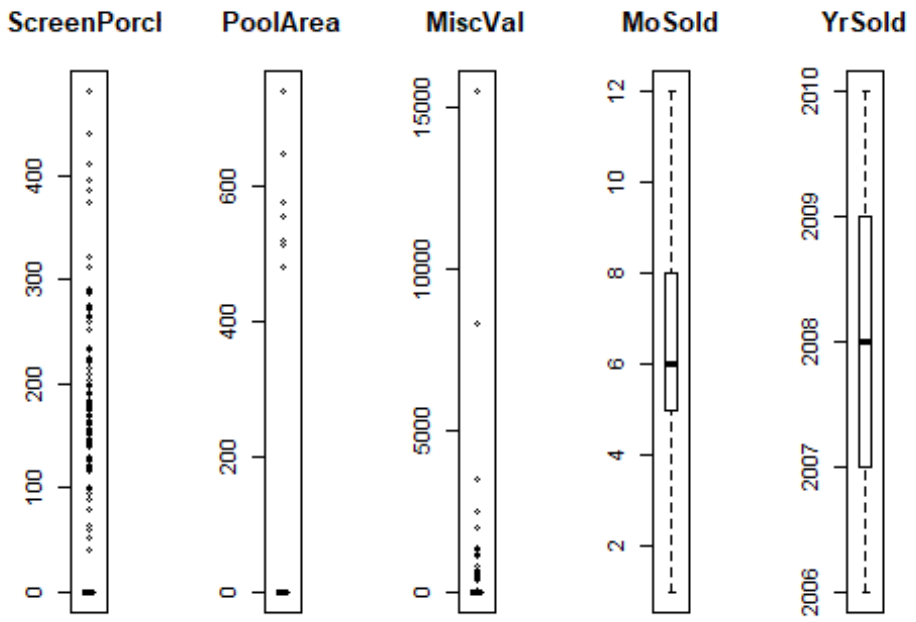

LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAd
MasVnrArea
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF


TotalBsmtSF
X1stFlrSF
X2ndFlrSF
LowQualFinS
GrLivArea
0
1000
2000
3000
4000
5000
6000
500
1500
100
200
300
400


BsmtFullBat
BsmtHalfBat
FullBath
HalfBath
BedroomAbv
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
2
4
6
8

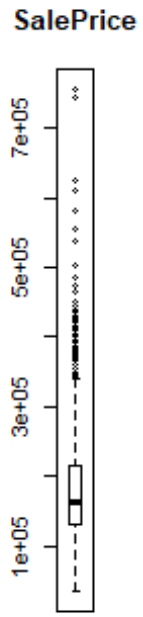

KitchenAbvG
TotRmsAbvGr
Fireplaces
GarageYrBlt
GarageCars
0.0 0.5 1.0 1.5 2.0 2.5 3.0
2 4 6 8 10 12 14
0.0 0.5 1.0 1.5 2.0 2.5 3.0
1900 1920 1940 1960 1980 2000
0 1 2 3 4


GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorc
X3SsnPorch
0 200 400 600 800 1000 1200 1400
0 200 400 600 800
0 100 200 300 400 500
0 100 200 300 400 500
0 100 200 300 400 500




```
##       LotFrontage LotArea    OverallQual OverallCond YearBuilt
## stats Numeric,5   Numeric,5  Numeric,5   Numeric,5   Numeric,5
## n     1201        1460       1460        1460        1460
## conf  Numeric,2   Numeric,2  Numeric,2   Numeric,2   Numeric,2
## out   Numeric,88  Numeric,68 Numeric,2   Numeric,125 Numeric,7
## group Numeric,88  Numeric,68 Numeric,2   Numeric,125 Numeric,7
## names ""          ""         ""          ""          ""
##       YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2  BsmtUnfSF
## stats Numeric,5    Numeric,5  Numeric,5  Numeric,5   Numeric,5
## n     1460         1452       1460       1460        1460
## conf  Numeric,2    Numeric,2  Numeric,2  Numeric,2   Numeric,2
## out   Numeric,0    Numeric,96 Numeric,7  Numeric,167 Numeric,29
## group Numeric,0    Numeric,96 Numeric,7  Numeric,167 Numeric,29
## names ""           ""         ""         ""          ""
##       TotalBsmtSF X1stFlrSF  X2ndFlrSF LowQualFinSF GrLivArea
## stats Numeric,5   Numeric,5  Numeric,5 Numeric,5    Numeric,5
## n     1460        1460       1460      1460         1460
## conf  Numeric,2   Numeric,2  Numeric,2 Numeric,2    Numeric,2
## out   Numeric,61  Numeric,20 Numeric,2 Numeric,26   Numeric,31
## group Numeric,61  Numeric,20 Numeric,2 Numeric,26   Numeric,31
## names ""          ""         ""        ""           ""
##       BsmtFullBath BsmtHalfBath FullBath  HalfBath  BedroomAbvGr
## stats Numeric,5    Numeric,5    Numeric,5 Numeric,5 Numeric,5
## n     1460         1460         1460      1460      1460
## conf  Numeric,2    Numeric,2    Numeric,2 Numeric,2 Numeric,2
## out   3            Numeric,82   Numeric,0 Numeric,0 Numeric,35
## group 1            Numeric,82   Numeric,0 Numeric,0 Numeric,35
```

```
## names ""           ""           ""         ""         ""
##       KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars
## stats Numeric,5    Numeric,5    Numeric,5  Numeric,5   Numeric,5
## n     1460         1460         1460       1379        1460
## conf  Numeric,2    Numeric,2    Numeric,2  Numeric,2   Numeric,2
## out   Numeric,68   Numeric,30   Numeric,5  Numeric,0   Numeric,5
## group Numeric,68   Numeric,30   Numeric,5  Numeric,0   Numeric,5
## names ""           ""           ""         ""          ""
##       GarageArea WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch
## stats Numeric,5  Numeric,5  Numeric,5   Numeric,5     Numeric,5
## n     1460       1460       1460        1460          1460
## conf  Numeric,2  Numeric,2  Numeric,2   Numeric,2     Numeric,2
## out   Numeric,21 Numeric,32 Numeric,77  Numeric,208   Numeric,24
## group Numeric,21 Numeric,32 Numeric,77  Numeric,208   Numeric,24
## names ""         ""         ""          ""            ""
##       ScreenPorch PoolArea  MiscVal    MoSold    YrSold    SalePrice
## stats Numeric,5   Numeric,5 Numeric,5  Numeric,5 Numeric,5 Numeric,5
## n     1460        1460      1460       1460      1460      1460
## conf  Numeric,2   Numeric,2 Numeric,2  Numeric,2 Numeric,2 Numeric,2
## out   Numeric,116 Numeric,7 Numeric,52 Numeric,0 Numeric,0 Numeric,61
## group Numeric,116 Numeric,7 Numeric,52 Numeric,0 Numeric,0 Numeric,61
## names ""          ""        ""         ""        ""        ""
```



Print Outliers

```
#boxplot.stats(housedata_num[,1])
for (i in 1:36)
```

```
{
(a<-boxplot.stats(housedata_num[,i]))

print(colnames(housedata_num)[i])
print(a$out)}
```

```
## [1] "LotFrontage"
##  [1] 112 115  24  21 121 122  24 120 134 141  24  24 174  21  21 174  21
## [18]  21 120 129 140 120 118 116 150  21 130  21  24  21 137  21  21  24
## [35] 130  24  21  21  21 120  24  24 144 114  24  21 128 116 149  21 313
## [52]  24  24  24 122 130 121  21 115  21  21  21 120  24  24  24 114 168
## [69] 182 134  24 120 118 138 160  24 152  21 124  21 313  24 153 120 129
## [86] 124  21  21
## [1] "LotArea"
##  [1]  50271  19900  21000  21453  19378  31770  22950  25419 159000  19296
## [11]  39104  19138  18386 215245 164660  20431  18800  53107  34650  22420
## [21]  21750  70761  53227  40094  32668  21872  21780  25095  46589  20896
## [31]  18450  21535  26178 115149  21695  53504  21384  28698  45600  17920
## [41]  25286  27650  24090  25000   1300  21286  21750  29959  18000  23257
## [51]  17755  35760  18030  35133  32463  18890  24682  23595  17871  36500
## [61]  63887  20781  25339  57200  20544  19690  21930  26142
## [1] "OverallQual"
## [1] 1 1
## [1] "OverallCond"
##   [1] 8 8 8 2 3 8 8 8 8 9 8 9 9 3 9 8 2 8 8 9 8 3 8 8 8 1 2 3 3 8 2 8 3 8
8
##  [36] 8 8 9 8 8 8 8 9 9 3 8 3 8 9 8 8 8 3 3 3 2 8 8 9 8 8 9 8 9 3 8 8 8 8
8
##  [71] 8 8 8 8 8 3 8 3 8 8 8 9 8 9 8 3 8 3 8 3 3 8 3 8 8 3 9 3 8 3 9 8 8 8
8
## [106] 8 8 8 9 3 8 8 8 8 9 8 3 9 9 3 8 8 8 9 9
## [1] "YearBuilt"
## [1] 1880 1880 1880 1882 1880 1875 1872
## [1] "YearRemodAdd"
## integer(0)
## [1] "MasVnrArea"
##  [1]  640  650  456 1031  573 1115  576  443  468  600  768  480 1129  436
## [15]  456  664  653  491  748  456  922  506  604  472  481 1600  616  870
## [29]  530  500  510  650  432  473  772  435  562  921  762  594  479  584
## [43]  420  459  452  513  472  660  528  464 1170  630  466  651  442  894
## [57]  513  673  603  860  424 1047  442  816  760  541  423  424  975  450
## [71]  423  571  480  425  660 1378  456  425  420  766  554  632  567  451
## [85]  621  788  796  428  564  579  705  731  420  448  426  438
## [1] "BsmtFinSF1"
## [1] 1810 1880 1904 2260 2188 2096 5644
## [1] "BsmtFinSF2"
##   [1]   32  668  486   93  491  506  712  362   41  169  869  150  670
28
##  [15] 1080  181  768  215  374  208  441  184  279  306  180  712  580
690
```

```
##  [29]  692  228  125 1063  620  175  820 1474  264  479  147  232  380
544
##  [43]  294  258  121  180  391  531  344  539  713  210  311 1120  165
532
##  [57]  279   96  495  180  174 1127  139  202  645  123  551  219  606
147
##  [71]  612  480  182  132  336  468  287   35  499  180  180  723  119
182
##  [85]   40  551  117  239   80  472   64 1057  127  630  480  128  377
764
##  [99]  345  539 1085  435  823  500  290  324  634  411  841 1061   93
466
## [113]  396  354  294  149  193  117  273  465  400  468   41  682   64
557
## [127]  230  106  791  240  287  547  391  469  177  108  374  600  492
211
## [141]  168   96 1031  438  375  144   81  906  608  276  661   68  173
972
## [155]  105  420  469  546  334  352  872  374  110  627  163 1029  290
## [1] "BsmtUnfSF"
##  [1] 1777 1768 1907 1686 2336 1694 2121 1869 2153 1969 1709 2042 1774 2046
## [15] 1836 1935 1926 1734 1800 1753 1905 1800 1710 1752 1694 1689 2002 1753
## [29] 1795
## [1] "TotalBsmtSF"
##  [1]    0    0 2223    0    0    0 2216    0 2392    0 2121 2136 3206    0
## [15]    0    0    0 3094 2153 3200    0 3138    0    0    0    0 2109 2077
## [29] 2444    0    0    0    0 2078    0 2217    0    0 2330    0    0    0
## [43]    0 2524    0    0    0    0    0 2396 2158    0    0 2136    0 2076
## [57] 2110 6110    0 2633    0
## [1] "X1stFlrSF"
##  [1] 2207 2223 2259 2158 2234 2392 2402 3228 3138 2515 2444 2217 2364 2898
## [15] 2524 2411 2196 4692 2156 2633
## [1] "X2ndFlrSF"
## [1] 1872 2065
## [1] "LowQualFinSF"
##  [1] 360 513 234 528 572 144 392 371 390 420 473 156 515 360  80  80  53
## [18] 232 481 120 514 397 479 205  80 384
## [1] "GrLivArea"
##  [1] 2945 3222 3608 3112 2794 3493 2978 3228 4676 2775 3194 3395 4316 3279
## [15] 3140 2822 2872 2898 3082 2868 2828 3627 3086 2872 4476 3447 5642 2810
## [29] 2792 3238 2784
## [1] "BsmtFullBath"
## [1] 3
## [1] "BsmtHalfBath"
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [36] 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1
## [71] 1 1 1 1 1 1 1 1 1 1 1
## [1] "FullBath"
## integer(0)
## [1] "HalfBath"
```

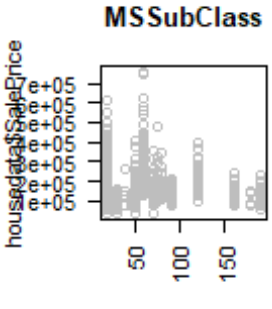
```
## integer(0)
## [1] "BedroomAbvGr"
##  [1] 0 5 5 6 0 5 6 5 5 6 5 6 5 0 8 5 6 5 5 6 5 5 5 5 5 5 5 0 0 5 0 5 6 5 5
## [1] "KitchenAbvGr"
##  [1] 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2
## [36] 2 2 2 2 2 2 2 2 2 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [1] "TotRmsAbvGrd"
##  [1] 11 11 12 11 11 11 11 14 11 12 11 12 11 11 12 11 12 11 12 11 11 12 12
## [24] 11 11 12 12 12 11 11
## [1] "Fireplaces"
## [1] 3 3 3 3 3
## [1] "GarageYrBlt"
## integer(0)
## [1] "GarageCars"
## [1] 4 4 4 4 4
## [1] "GarageArea"
##  [1] 1166  968 1053 1025  947 1390 1134  983 1020 1220 1248 1043 1052  995
## [15] 1356 1052  954 1014 1418  968 1069
## [1] "WoodDeckSF"
##  [1] 857 576 476 574 441 468 670 495 536 519 466 517 426 503 486 486 511
## [18] 421 550 509 474 728 436 431 448 439 635 500 668 586 431 736
## [1] "OpenPorchSF"
##  [1] 204 213 258 199 234 184 205 228 238 260 198 172 208 228 184 250 175
## [18] 195 214 231 192 187 176 523 285 406 182 502 274 172 243 235 312 267
## [35] 265 288 341 204 174 247 291 312 418 240 364 188 207 234 192 191 252
## [52] 189 282 224 319 244 185 200 180 263 304 234 240 192 229 211 198 287
## [69] 292 207 241 547 211 184 262 210 236
## [1] "EnclosedPorch"
##   [1] 272 228 205 176 205  87 172 102  37 144  64 114 202 128 156  44  77
##  [18] 144 192 144 140 180 228 128 183  39 184  40 552  30 126  96  60 150
##  [35] 120 202  77 112 252  52 224 234 144 244 268 137  24 108 294 177 218
##  [52] 242  91 112 160 130 184 126 169 105  34  96 248 236 120  32  80 115
##  [69] 291 184 116 158 112 210  36 156 144 200  84 148 116 120 136 102 240
##  [86]  54 112  39 100  36 189 293 164  40 216 239 112 252 240 180  67  90
## [103] 120  56 112 129  40  98 143 216 234 112 112  70 386 154 185 156 156
## [120] 134 196 264 185 275  96 120 112 116 230 254  68 194 192  34 150 164
## [137] 112 224  32 318 244  48  94 138 108 112 226 192 174 228  19 170 220
## [154] 128  80 115 137 192 252 112  96 176 216 176 214 280  96 116 102 190
## [171] 236 192  84 330 208 145 259 126 264  81 164  42 123 162 100 286 190
## [188] 168  20 301 198  96 221 112 212  50 150 168 112 160 114 216 154  99
## [205] 158 216 252 112
## [1] "X3SsnPorch"
##  [1] 320 407 130 180 168 180 140 508 238 245 196 144 144 182 168 162  23
## [18] 168 216  96 216 153 290 304
## [1] "ScreenPorch"
##   [1] 176 198 291 252  99 184 168 130 142 192 410 224 266 170 154 153 144
##  [18] 142 128 259 160 198 271 234 184 374 192 185 182  90 144 224 396 170
##  [35] 176 140 276 192 180 161 168 145 200 122  95 144 120  60 120 126 189
##  [52] 260 147 385 287 200 156 100 180 216 210 197 204 192 225 192 152 175
##  [69] 126 312 222 265 224 322 120 190 233  63 147 180  53 143 189 189 189
```

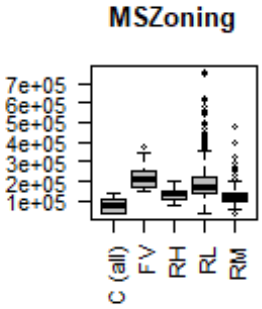
```
##  [86] 192 160 160 126 100 273 180  90 288 263 224 147 120  80 163  90 288
## [103] 116 259 224 216 480 120 178 440 155 168 220 119 165  40
## [1] "PoolArea"
## [1] 512 648 576 555 480 519 738
## [1] "MiscVal"
##  [1]   700   350   700   500   400   700   480   400   400   450   450
## [12]   500   450   700   400 15500  1200   800   480   400  2000  2000
## [23]   600   500   600   600  3500   500   400   450   500  1300  1200
## [34]   500   400    54   500   400   400  2000   620   400   560   500
## [45]   700  1400   400  8300   600  1150  2000  2500
## [1] "MoSold"
## integer(0)
## [1] "YrSold"
## integer(0)
## [1] "SalePrice"
##  [1] 345000 385000 438780 383970 372402 412500 501837 475000 386250 403000
## [11] 415298 360000 375000 342643 354000 377426 437154 394432 426000 555000
## [21] 440000 380000 374000 430000 402861 446261 369900 451950 359100 345000
## [31] 370878 350000 402000 423000 372500 392000 755000 361919 341000 538000
## [41] 395000 485000 582933 385000 350000 611657 395192 348000 556581 424870
## [51] 625000 392500 745000 367294 465000 378500 381000 410000 466500 377500
## [61] 394617
```

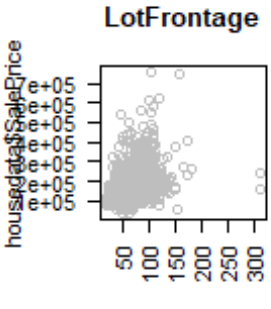
Looking through these values, most of them looks to be biproduct of skewed distribution or extremely small variance. Will look this into a little more after cleaning up the data using cook's distance.

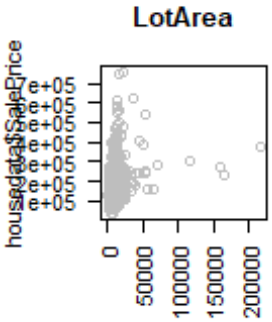
Check to see relationship with response variable - house price VS ind. variables.

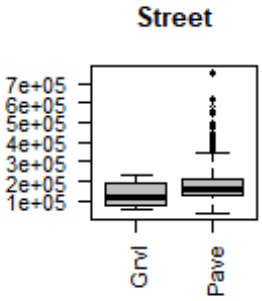
```
par(mfrow=c(2,3))
for (i in 2:79)
(
plot(housedata[,i],housedata$SalePrice, las=2,  main=colnames(housedata[i]),
xlab='', col='grey')
)
```

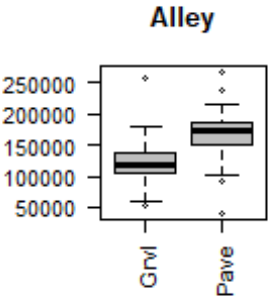

MSSubClass
housedata1$SalePrice
50
100
150

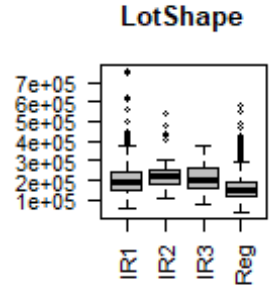

MSZoning
C (all)
FV
RH
RL
RM

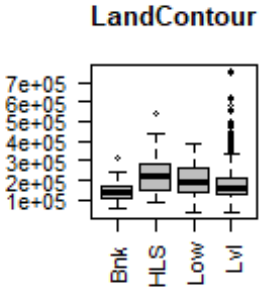

LotFrontage
housedata1$SalePrice
50
100
150
200
250
300

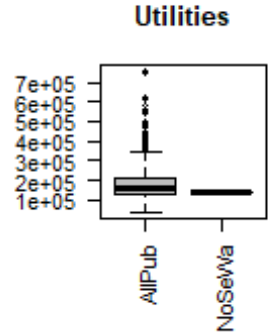

LotArea
housedata1$SalePrice
0
50000
100000
150000
200000

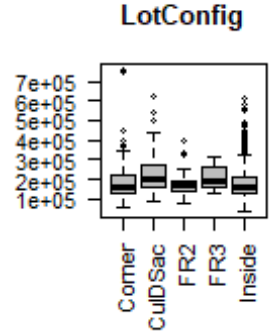

Street
Grvl
Pave

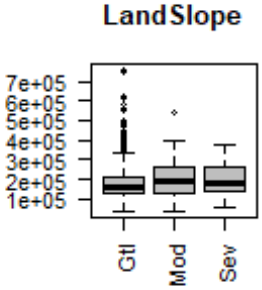

Alley
Grvl
Pave

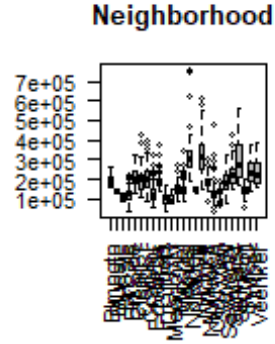

LotShape
IR1
IR2
IR3
Reg

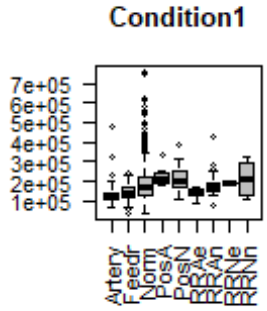

LandContour
Bnk
HLS
Low
Lvl

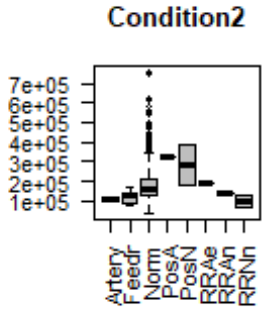

Utilities
AllPub
NoSeWa

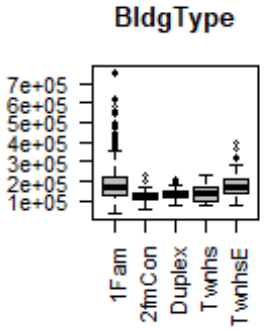

LotConfig
Corner
CulDSac
FR2
FR3
Inside

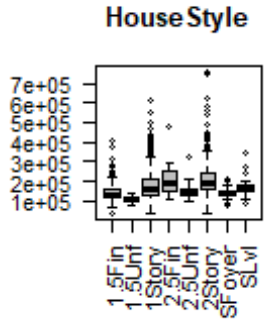

LandSlope
Gtl
Mod
Sev

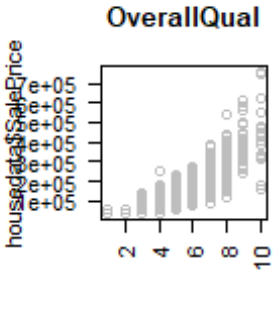

Neighborhood

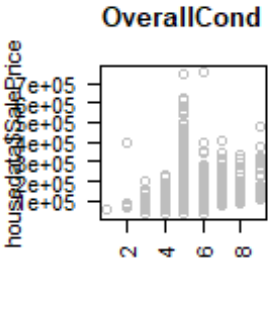

Condition1
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
Artery
Feedr
Norm
PosA
PosN
RRAe
RRAn
RRNe
RRNn

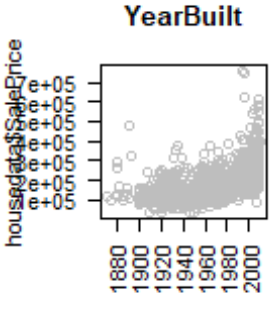

Condition2
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
Artery
Feedr
Norm
PosA
PosN
RRAe
RRAn
RRNn

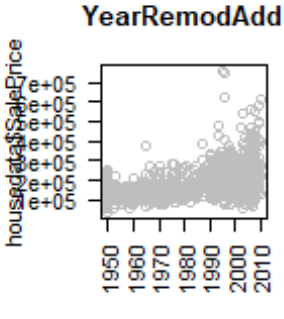

BldgType
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
1Fam
2fmCon
Duplex
Twnhs
TwnhsE

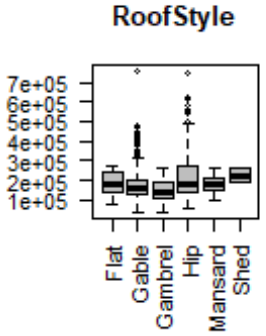

HouseStyle
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
1.5Fin
1.5Unf
1Story
2.5Fin
2.5Unf
2Story
SFoyer
SLvl

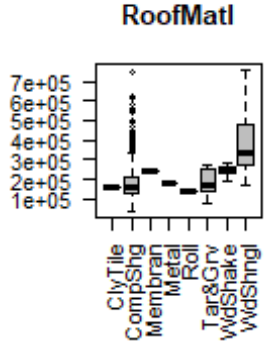

OverallQual
housedata$SalePrice
2
4
6
8
10

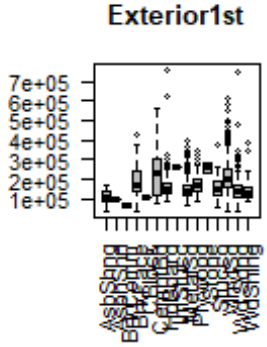

OverallCond
housedata$SalePrice
2
4
6
8

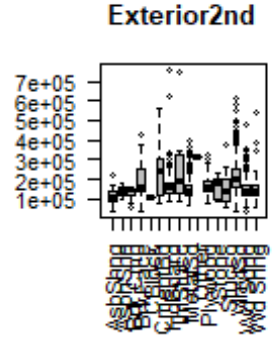

YearBuilt
housedata$SalePrice
1880
1900
1920
1940
1960
1980
2000

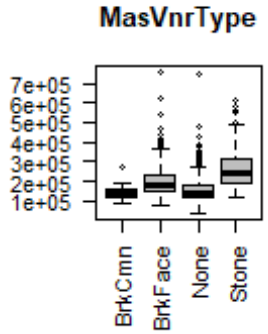

YearRemodAdd
housedata$SalePrice
1950
1960
1970
1980
1990
2000
2010

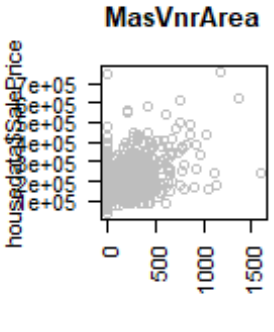

RoofStyle
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
Flat
Gable
Gambrel
Hip
Mansard
Shed

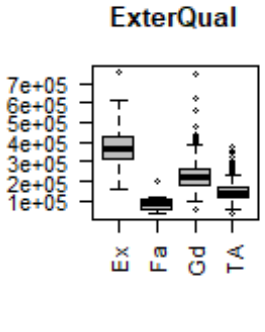

RoofMatl
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
ClyTile
CompShg
Membran
Metal
Roll
Tar&Grv
WdShake
WdShngl

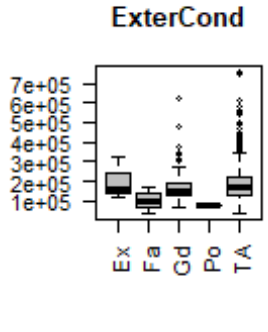

Exterior1st
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05

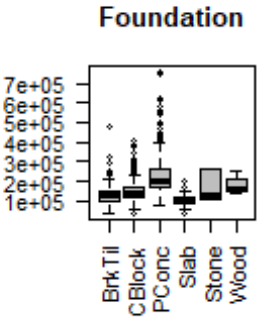

Exterior2nd
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05

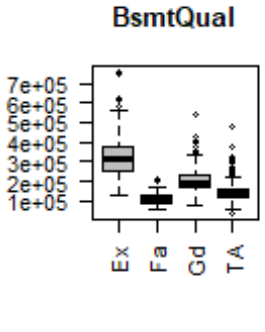

MasVnrType
BrkCmn
BrkFace
None
Stone

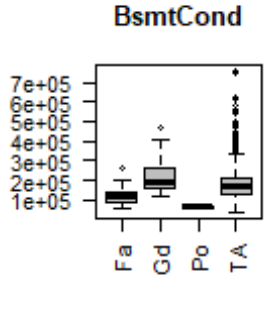

MasVnrArea
housedata$SalePrice
0
500
1000
1500

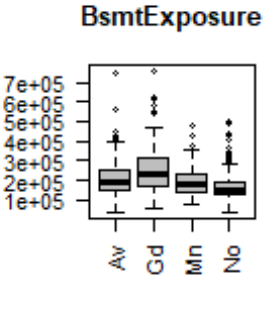

ExterQual
Ex
Fa
Gd
TA

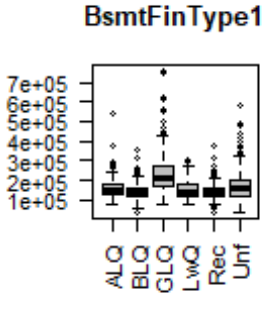

ExterCond
Ex
Fa
Gd
Po
TA

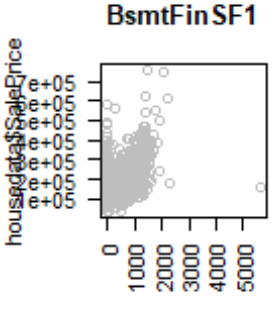

Foundation
BrkTil
CBlock
PConc
Slab
Stone
Wood

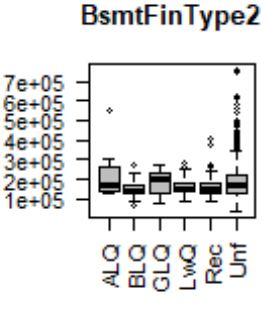

BsmtQual
Ex
Fa
Gd
TA

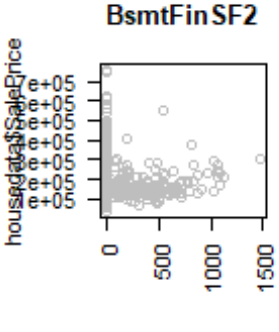

BsmtCond
Fa
Gd
Po
TA

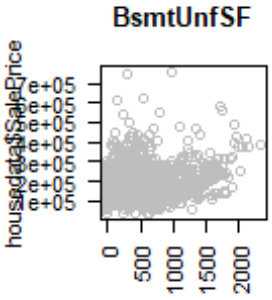

BsmtExposure
Av
Gd
Mn
No

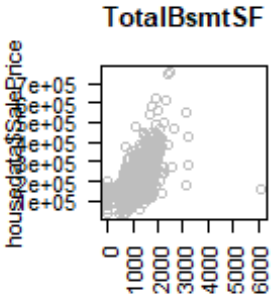

BsmtFinType1
ALQ
BLQ
GLQ
LwQ
Rec
Unf

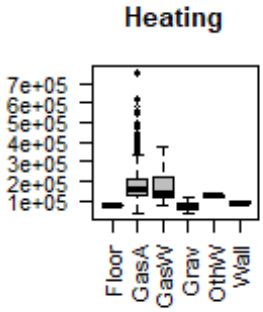

BsmtFinSF1
housedata$SalePrice
0
1000
2000
3000
4000
5000

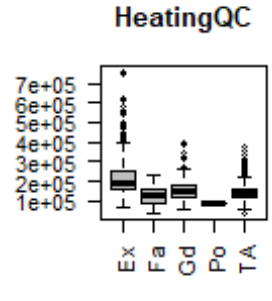

BsmtFinType2
ALQ
BLQ
GLQ
LwQ
Rec
Unf

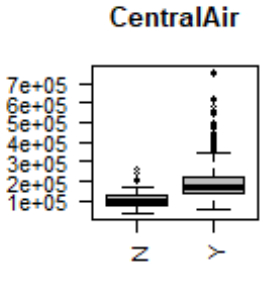

BsmtFinSF2
housedata$SalePrice
0
500
1000
1500

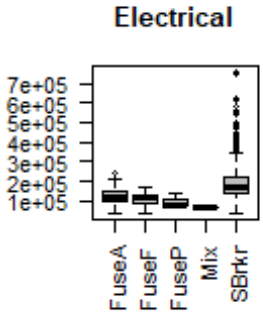

BsmtUnfSF
housedata$SalePrice
0
500
1000
1500
2000

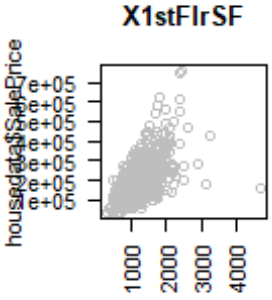

TotalBsmtSF
housedata$SalePrice
0
1000
2000
3000
4000
5000
6000

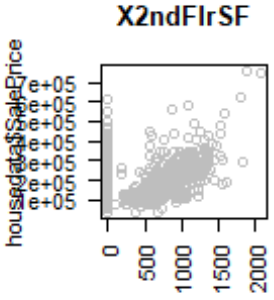

Heating
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
Floor
GasA
GasW
Grav
OthW
Wall

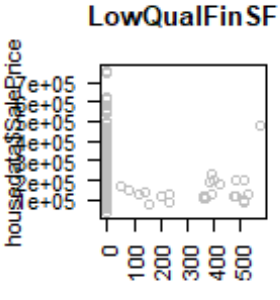

HeatingQC
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
Ex
Fa
Gd
Po
TA

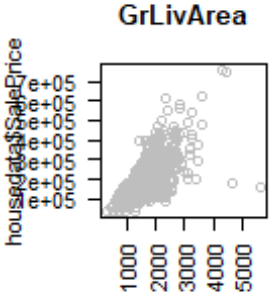

CentralAir
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
N
Y

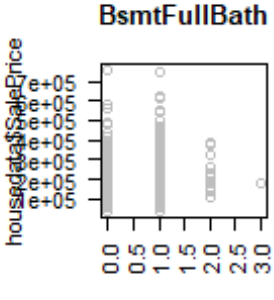

Electrical
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
FuseA
FuseF
FuseP
Mix
SBrkr

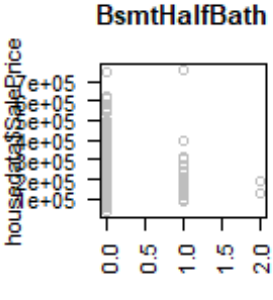

X1stFlrSF
housedata$SalePrice
1000
2000
3000
4000

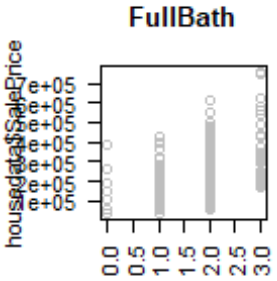

X2ndFlrSF
housedata$SalePrice
0
500
1000
1500
2000

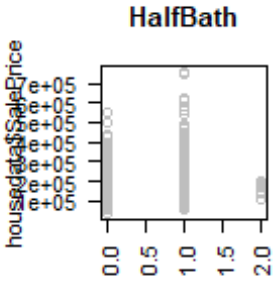

LowQualFinSF
housedata$SalePrice
0
100
200
300
400
500

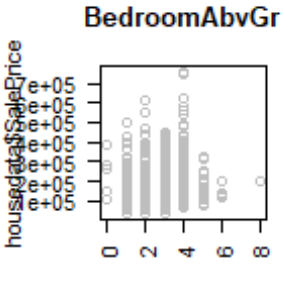

GrLivArea
housedata$SalePrice
1000
2000
3000
4000
5000

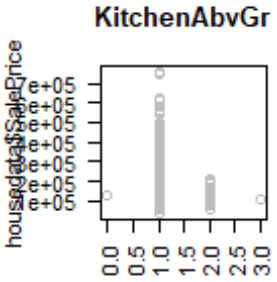

BsmtFullBath
housedata$SalePrice
0.0
0.5
1.0
1.5
2.0
2.5
3.0

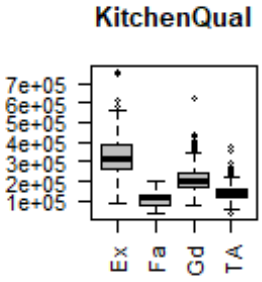

BsmtHalfBath
housedata$SalePrice
0.0
0.5
1.0
1.5
2.0

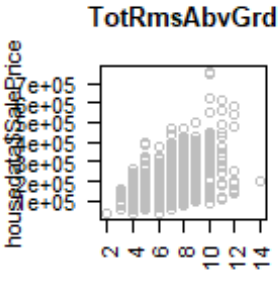

FullBath
housedata$SalePrice
0.0
0.5
1.0
1.5
2.0
2.5
3.0

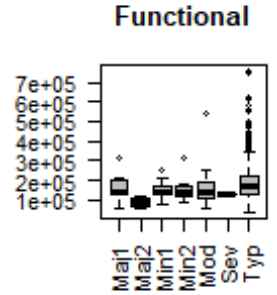

HalfBath
housedata$SalePrice
0.0
0.5
1.0
1.5
2.0

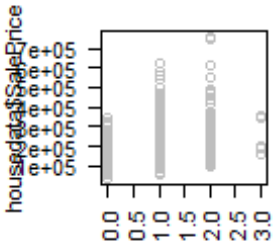

BedroomAbvGr
housedata$SalePrice
0
2
4
6
8

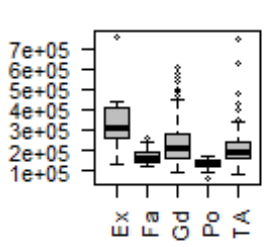

KitchenAbvGr
housedata$SalePrice
0.0
0.5
1.0
1.5
2.0
2.5
3.0

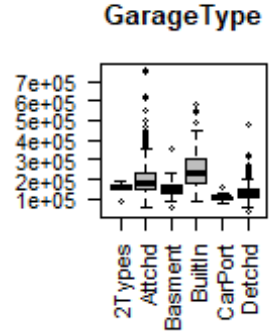

KitchenQual
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
Ex
Fa
Gd
TA

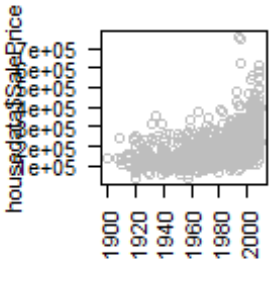

TotRmsAbvGrd
housedata$SalePrice
2
4
6
8
10
12
14

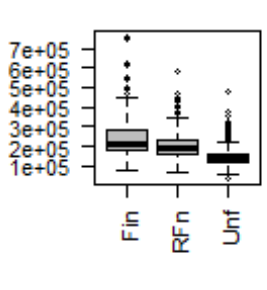

Functional
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
Maj1
Maj2
Min1
Min2
Mod
Sev
Typ

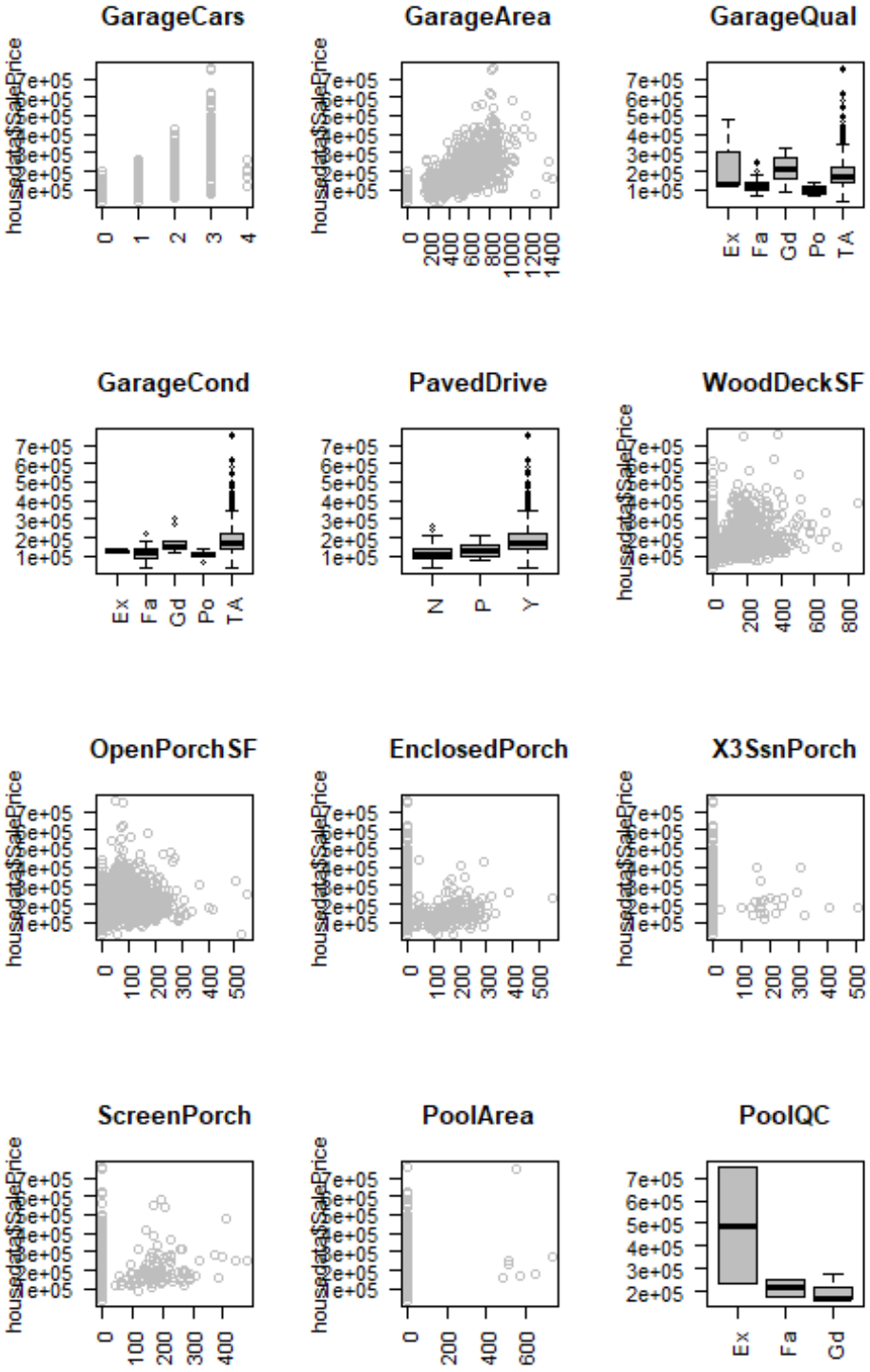

Fireplaces
housedata$SalePrice
0.0
0.5
1.0
1.5
2.0
2.5
3.0

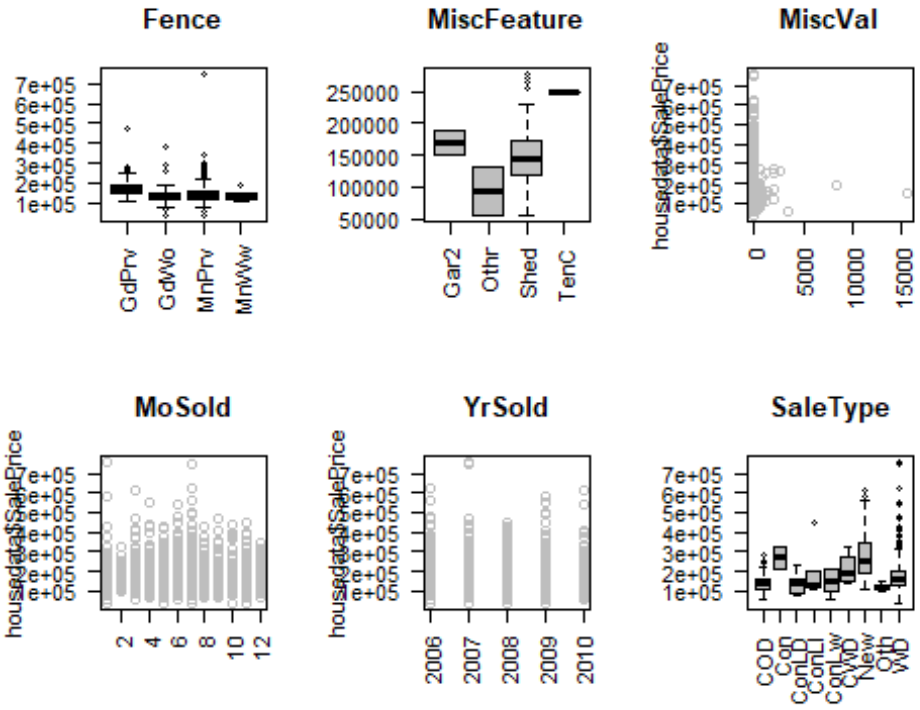

FireplaceQu
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
Ex
Fa
Gd
Po
TA


GarageType
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
2Types
Attchd
Basment
BuiltIn
CarPort
Detchd


GarageYrBlt
housedata$SalePrice
1900
1920
1940
1960
1980
2000


GarageFinish
7e+05
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
Fin
RFn
Unf


GarageCars
GarageArea
GarageQual
GarageCond
PavedDrive
WoodDeckSF
OpenPorchSF
EnclosedPorch
X3SsnPorch
ScreenPorch
PoolArea
PoolQC



Can see definate relationships here. House Prices increase with Lot Frontage, Lot Area, Overall Quality, Year Built, Year Remodel, BsmtFinSF1, TotalBsmtSF, X1stFlrSF, X2ndFlrSF, # of FullBath, TotRmsAbvGrd,KitchenQual, GarageYrBlt, GarageCars, GarageArea, WoodDeckSF, etc.

Checking Correlations between Independent Variables

Using spearman because many of our variables are not normal.

```
cor(housedata_num,use="complete.obs", method='spearman')
```

```
##                 LotFrontage     LotArea OverallQual  OverallCond
## LotFrontage    1.0000000000  0.64436154  0.24008405 -0.069072924
## LotArea        0.6443615430  1.00000000  0.27885345 -0.088291269
## OverallQual    0.2400840531  0.27885345  1.00000000 -0.234850009
## OverallCond   -0.0690729241 -0.08829127 -0.23485001  1.000000000
## YearBuilt      0.1749910491  0.13128094  0.66393187 -0.453285876
## YearRemodAdd   0.1125724029  0.12215453  0.59165419 -0.097301007
## MasVnrArea     0.2548257439  0.17379506  0.42994108 -0.218477063
## BsmtFinSF1     0.1561913951  0.14435735  0.12571321 -0.012018369
## BsmtFinSF2     0.0507176625  0.02144473 -0.13324890  0.110054962
## BsmtUnfSF      0.1026648516  0.09935269  0.29049024 -0.144815183
## TotalBsmtSF    0.3711624304  0.36042189  0.50014015 -0.243558659
## X1stFlrSF      0.4101593659  0.41960646  0.44944825 -0.189691784
## X2ndFlrSF      0.0524942879  0.13632122  0.26104972 -0.016963283
## LowQualFinSF  -0.0505368659 -0.02056586 -0.02632446  0.034550601
## GrLivArea      0.3623017426  0.45698925  0.62075325 -0.184906470
```

```
## BsmtFullBath   0.0957102680  0.10049548  0.11185274 -0.056597923
## BsmtHalfBath  -0.0009884616 -0.01080862 -0.06528073  0.126624347
## FullBath       0.2079144047  0.23270443  0.60532881 -0.295229283
## HalfBath       0.0867115850  0.15379215  0.26908590 -0.083215943
## BedroomAbvGr   0.3276698150  0.33715069  0.12169997 -0.002472406
## KitchenAbvGr  -0.0059603190 -0.02968873 -0.18258511 -0.105298357
## TotRmsAbvGrd   0.3610374056  0.42633210  0.45022129 -0.135635588
## Fireplaces     0.2350587665  0.33789026  0.43316874 -0.041139036
## GarageYrBlt    0.1149244614  0.11010050  0.63266717 -0.401291480
## GarageCars     0.3413303267  0.34109772  0.59364757 -0.316159374
## GarageArea     0.3703701038  0.35599905  0.53027865 -0.260390407
## WoodDeckSF     0.1020957595  0.16802967  0.30227528 -0.052770647
## OpenPorchSF    0.1793140525  0.18991125  0.45885277 -0.200632164
## EnclosedPorch -0.0976292472 -0.07993287 -0.20052338  0.110328966
## X3SsnPorch     0.0623898717  0.04470101  0.01547530  0.009856910
## ScreenPorch    0.0367199004  0.10682888  0.03292948  0.107517740
## PoolArea       0.0860921624  0.08685056  0.06578001 -0.023172667
## MiscVal        0.0207394372  0.04701516 -0.10732995  0.108376809
## MoSold         0.0305732569  0.01173592  0.07992739 -0.015541322
## YrSold         0.0076572524 -0.02530713 -0.01030941  0.041029181
## SalePrice      0.3996569514  0.46028741  0.82710824 -0.174954934
##                   YearBuilt YearRemodAdd   MasVnrArea   BsmtFinSF1
## LotFrontage    0.1749910491  0.112572403  0.254825744  0.156191395
## LotArea        0.1312809411  0.122154528  0.173795062  0.144357349
## OverallQual    0.6639318684  0.591654188  0.429941084  0.125713206
## OverallCond   -0.4532858756 -0.097301007 -0.218477063 -0.012018369
## YearBuilt      1.0000000000  0.730853110  0.425485771  0.165970844
## YearRemodAdd   0.7308531105  1.000000000  0.262924835  0.049455279
## MasVnrArea     0.4254857705  0.262924835  1.000000000  0.236420516
## BsmtFinSF1     0.1659708443  0.049455279  0.236420516  1.000000000
## BsmtFinSF2    -0.1222026052 -0.126092469 -0.071561654  0.065509767
## BsmtUnfSF      0.1717265356  0.195982824  0.082558734 -0.595562782
## TotalBsmtSF    0.4553712555  0.342818103  0.389679664  0.391192733
## X1stFlrSF      0.3307112127  0.290261699  0.377821545  0.324099436
## X2ndFlrSF     -0.0003047781  0.033900558  0.054817554 -0.189395375
## LowQualFinSF  -0.1182972945 -0.049010374 -0.099228290 -0.067831402
## GrLivArea      0.2990370074  0.293263578  0.344006036  0.059425782
## BsmtFullBath   0.1515273868  0.093558480  0.143593672  0.674398591
## BsmtHalfBath  -0.0752161601 -0.045372860 -0.006664976  0.090677191
## FullBath       0.5677105014  0.467212486  0.306071781  0.001472189
## HalfBath       0.2120053123  0.129045365  0.149984263 -0.023180761
## BedroomAbvGr  -0.0369216046 -0.080779607  0.109417542 -0.083442810
## KitchenAbvGr  -0.1509401322 -0.178517673 -0.036494620 -0.088603536
## TotRmsAbvGrd   0.2040950314  0.197404528  0.296649476 -0.029543442
## Fireplaces     0.1668439232  0.139463911  0.253584164  0.188202808
## GarageYrBlt    0.8927077331  0.740592061  0.348486318  0.082764465
## GarageCars     0.6011012262  0.492888530  0.396932750  0.133503072
## GarageArea     0.5380341421  0.439150316  0.364521183  0.200467636
## WoodDeckSF     0.2994071607  0.261491485  0.204879535  0.174923779
## OpenPorchSF    0.4421110389  0.384326792  0.230627560  0.086836515
```

```
## EnclosedPorch -0.4153538745 -0.267526111 -0.196440656 -0.151555666
## X3SsnPorch     0.0193667880  0.030931795  0.044434329  0.036169664
## ScreenPorch   -0.0799940411 -0.038799487  0.033942816  0.078383880
## PoolArea       0.0105691057  0.009272607  0.014386412  0.086986654
## MiscVal       -0.1173028153 -0.100067923 -0.094631831  0.010187895
## MoSold         0.0213327029  0.025676337  0.029048504 -0.008925338
## YrSold        -0.0046510171  0.050791355 -0.011500579  0.014983327
## SalePrice      0.6701983700  0.617311189  0.435443677  0.285329793
##                  BsmtFinSF2    BsmtUnfSF  TotalBsmtSF    X1stFlrSF
## LotFrontage    0.050717663  0.102664852  0.371162430  0.410159366
## LotArea        0.021444732  0.099352687  0.360421891  0.419606462
## OverallQual   -0.133248897  0.290490242  0.500140154  0.449448245
## OverallCond    0.110054962 -0.144815183 -0.243558659 -0.189691784
## YearBuilt     -0.122202605  0.171726536  0.455371255  0.330711213
## YearRemodAdd  -0.126092469  0.195982824  0.342818103  0.290261699
## MasVnrArea    -0.071561654  0.082558734  0.389679664  0.377821545
## BsmtFinSF1     0.065509767 -0.595562782  0.391192733  0.324099436
## BsmtFinSF2     1.000000000 -0.294576597  0.041040989  0.029313173
## BsmtUnfSF     -0.294576597  1.000000000  0.338722939  0.227874470
## TotalBsmtSF    0.041040989  0.338722939  1.000000000  0.839877975
## X1stFlrSF      0.029313173  0.227874470  0.839877975  1.000000000
## X2ndFlrSF     -0.121873067  0.044758217 -0.302324596 -0.302256239
## LowQualFinSF   0.004304572  0.004235821 -0.089415311 -0.050472084
## GrLivArea     -0.097353408  0.242976508  0.370490663  0.477355952
## BsmtFullBath   0.159750855 -0.457770811  0.296196670  0.244145012
## BsmtHalfBath   0.091965341 -0.110764767 -0.018035495 -0.002438725
## FullBath      -0.124438016  0.268350504  0.345973505  0.368985245
## HalfBath      -0.057573583 -0.023604508 -0.126497595 -0.179842722
## BedroomAbvGr   0.007525140  0.132639841  0.046973860  0.115589535
## KitchenAbvGr  -0.058900663 -0.026200836 -0.041696014  0.086706711
## TotRmsAbvGrd  -0.100899621  0.229519580  0.236645419  0.352118505
## Fireplaces     0.002670659  0.071630353  0.328306925  0.391655841
## GarageYrBlt   -0.145341137  0.209607477  0.385349349  0.285489869
## GarageCars    -0.110580365  0.204513211  0.477102844  0.462872297
## GarageArea    -0.050814936  0.146350368  0.499281734  0.489323943
## WoodDeckSF     0.045529152 -0.021810468  0.238384210  0.226807371
## OpenPorchSF   -0.090767878  0.163910596  0.310282171  0.253944843
## EnclosedPorch  0.037317210  0.018886754 -0.195629093 -0.156584494
## X3SsnPorch    -0.024518933  0.022902517  0.048304585  0.050413909
## ScreenPorch    0.039760171 -0.014195470  0.079378295  0.101152573
## PoolArea       0.091224798 -0.056655367  0.070588661  0.070350280
## MiscVal        0.018661546 -0.042736789 -0.063806227 -0.047238000
## MoSold        -0.060286644  0.025457811  0.019426598  0.044991883
## YrSold         0.036695437 -0.027446064 -0.008978755 -0.015761290
## SalePrice     -0.076097825  0.201583146  0.618297886  0.582929810
##                   X2ndFlrSF  LowQualFinSF   GrLivArea BsmtFullBath
## LotFrontage    0.0524942879 -0.0505368659  0.36230174  0.095710268
## LotArea        0.1363212181 -0.0205658645  0.45698925  0.100495484
## OverallQual    0.2610497237 -0.0263244611  0.62075325  0.111852744
## OverallCond   -0.0169632826  0.0345506009 -0.18490647 -0.056597923
```

```
## YearBuilt     -0.0003047781 -0.1182972945  0.29903701  0.151527387
## YearRemodAdd   0.0339005576 -0.0490103736  0.29326358  0.093558480
## MasVnrArea     0.0548175541 -0.0992282901  0.34400604  0.143593672
## BsmtFinSF1    -0.1893953753 -0.0678314018  0.05942578  0.674398591
## BsmtFinSF2    -0.1218730670  0.0043045717 -0.09735341  0.159750855
## BsmtUnfSF      0.0447582169  0.0042358211  0.24297651 -0.457770811
## TotalBsmtSF   -0.3023245960 -0.0894153108  0.37049066  0.296196670
## X1stFlrSF     -0.3022562390 -0.0504720836  0.47735595  0.244145012
## X2ndFlrSF      1.0000000000  0.0486235981  0.63605365 -0.168995104
## LowQualFinSF   0.0486235981  1.0000000000  0.03941525 -0.042711209
## GrLivArea      0.6360536500  0.0394152519  1.00000000  0.027019207
## BsmtFullBath  -0.1689951043 -0.0427112092  0.02701921  1.000000000
## BsmtHalfBath  -0.0333066710  0.0025545202 -0.05303238 -0.141587489
## FullBath       0.3618525345 -0.0106333232  0.64526788 -0.039813488
## HalfBath       0.6253557214  0.0083894435  0.42059827 -0.039045569
## BedroomAbvGr   0.5110462045  0.0157780267  0.53191648 -0.097265496
## KitchenAbvGr   0.0487330406 -0.0259574290  0.11832523 -0.055074367
## TotRmsAbvGrd   0.5810123707  0.0122257280  0.82767979 -0.035088921
## Fireplaces     0.1877894772 -0.0195792220  0.48233093  0.109448583
## GarageYrBlt    0.0437806504 -0.0374235875  0.30111666  0.099306670
## GarageCars     0.1395414342 -0.0318513980  0.50952492  0.132195453
## GarageArea     0.0812539747 -0.0146065686  0.46539364  0.194232917
## WoodDeckSF     0.0901093829 -0.0455740252  0.24891999  0.151001396
## OpenPorchSF    0.2002104280  0.0124866618  0.40116818  0.104174804
## EnclosedPorch  0.0577403115  0.0166793036 -0.07155494 -0.069209765
## X3SsnPorch    -0.0265261776  0.0411341615  0.01636702 -0.004963405
## ScreenPorch    0.0172920968 -0.0090294951  0.08909547  0.030869612
## PoolArea       0.0664114701  0.0912017662  0.06522928  0.097319847
## MiscVal       -0.0299783732  0.0622955610 -0.07936888  0.016346189
## MoSold         0.0514682375 -0.0006644359  0.08515333 -0.031206461
## YrSold        -0.0217672068 -0.0192163798 -0.01601950  0.055227663
## SalePrice      0.2636923430 -0.0534196348  0.72497669  0.228001195
##                BsmtHalfBath     FullBath     HalfBath BedroomAbvGr
## LotFrontage   -0.0009884616  0.207914405  0.086711585  0.327669815
## LotArea       -0.0108086164  0.232704430  0.153792147  0.337150691
## OverallQual   -0.0652807325  0.605328808  0.269085905  0.121699971
## OverallCond    0.1266243475 -0.295229283 -0.083215943 -0.002472406
## YearBuilt     -0.0752161601  0.567710501  0.212005312 -0.036921605
## YearRemodAdd  -0.0453728599  0.467212486  0.129045365 -0.080779607
## MasVnrArea    -0.0066649764  0.306071781  0.149984263  0.109417542
## BsmtFinSF1     0.0906771913  0.001472189 -0.023180761 -0.083442810
## BsmtFinSF2     0.0919653406 -0.124438016 -0.057573583  0.007525140
## BsmtUnfSF     -0.1107647666  0.268350504 -0.023604508  0.132639841
## TotalBsmtSF   -0.0180354949  0.345973505 -0.126497595  0.046973860
## X1stFlrSF     -0.0024387252  0.368985245 -0.179842722  0.115589535
## X2ndFlrSF     -0.0333066710  0.361852535  0.625355721  0.511046204
## LowQualFinSF   0.0025545202 -0.010633323  0.008389444  0.015778027
## GrLivArea     -0.0530323820  0.645267881  0.420598273  0.531916481
## BsmtFullBath  -0.1415874891 -0.039813488 -0.039045569 -0.097265496
## BsmtHalfBath   1.0000000000 -0.082501598 -0.028061366  0.024842309
```

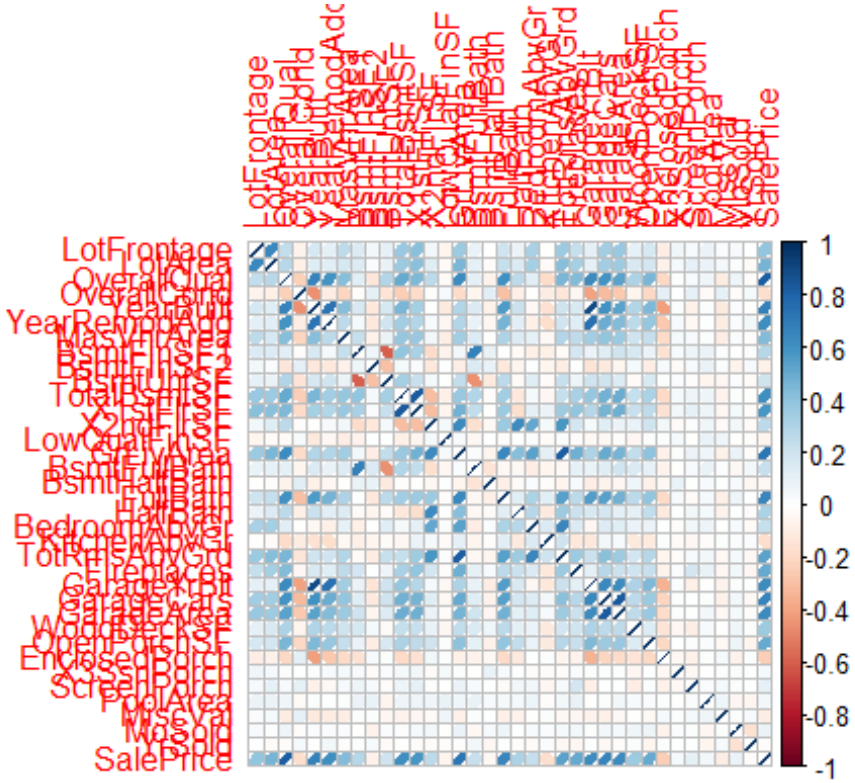
```
## FullBath      -0.0825015983  1.000000000  0.124288926  0.311891841
## HalfBath      -0.0280613662  0.124288926  1.000000000  0.274207599
## BedroomAbvGr   0.0248423091  0.311891841  0.274207599  1.000000000
## KitchenAbvGr  -0.0110093387  0.122242067 -0.115170146  0.150573097
## TotRmsAbvGrd  -0.0544288956  0.546450657  0.353408624  0.652004681
## Fireplaces     0.0323297696  0.268739961  0.212271790  0.159547811
## GarageYrBlt   -0.0903053650  0.555989021  0.166465199 -0.040597341
## GarageCars    -0.0616718084  0.545065468  0.188727249  0.142951469
## GarageArea    -0.0494521515  0.472821175  0.114172960  0.132319982
## WoodDeckSF     0.0624413721  0.260847826  0.112320643  0.084445141
## OpenPorchSF   -0.0786393834  0.394750044  0.246707280  0.081954835
## EnclosedPorch -0.0366503950 -0.196792395 -0.117676999  0.006239924
## X3SsnPorch     0.0940222919  0.036521668  0.002922579 -0.019547495
## ScreenPorch    0.0093612978 -0.026848815  0.058725619  0.051641963
## PoolArea       0.0361075137  0.040220228  0.043386207  0.070511913
## MiscVal        0.0031499295 -0.083238890 -0.062615754 -0.008164026
## MoSold         0.0332711266  0.079984189  0.002701294  0.034631945
## YrSold        -0.0475053838  0.008607973 -0.018867453 -0.026042899
## SalePrice     -0.0394004392  0.651157919  0.319917623  0.234986242
##               KitchenAbvGr TotRmsAbvGrd   Fireplaces  GarageYrBlt
## LotFrontage   -0.005960319   0.36103741  0.235058767  0.114924461
## LotArea       -0.029688728   0.42633210  0.337890257  0.110100499
## OverallQual   -0.182585115   0.45022129  0.433168740  0.632667167
## OverallCond   -0.105298357  -0.13563559 -0.041139036 -0.401291480
## YearBuilt     -0.150940132   0.20409503  0.166843923  0.892707733
## YearRemodAdd  -0.178517673   0.19740453  0.139463911  0.740592061
## MasVnrArea    -0.036494620   0.29664948  0.253584164  0.348486318
## BsmtFinSF1    -0.088603536  -0.02954344  0.188202808  0.082764465
## BsmtFinSF2    -0.058900663  -0.10089962  0.002670659 -0.145341137
## BsmtUnfSF     -0.026200836   0.22951958  0.071630353  0.209607477
## TotalBsmtSF   -0.041696014   0.23664542  0.328306925  0.385349349
## X1stFlrSF      0.086706711   0.35211850  0.391655841  0.285489869
## X2ndFlrSF      0.048733041   0.58101237  0.187789477  0.043780650
## LowQualFinSF  -0.025957429   0.01222573 -0.019579222 -0.037423588
## GrLivArea      0.118325230   0.82767979  0.482330931  0.301116657
## BsmtFullBath  -0.055074367  -0.03508892  0.109448583  0.099306670
## BsmtHalfBath  -0.011009339  -0.05442890  0.032329770 -0.090305365
## FullBath       0.122242067   0.54645066  0.268739961  0.555989021
## HalfBath      -0.115170146   0.35340862  0.212271790  0.166465199
## BedroomAbvGr   0.150573097   0.65200468  0.159547811 -0.040597341
## KitchenAbvGr   1.000000000   0.20032207 -0.123670258 -0.124936843
## TotRmsAbvGrd   0.200322068   1.00000000  0.364488649  0.219646027
## Fireplaces    -0.123670258   0.36448865  1.000000000  0.111077271
## GarageYrBlt   -0.124936843   0.21964603  0.111077271  1.000000000
## GarageCars     0.053060375   0.41807971  0.279547716  0.645409535
## GarageArea     0.037465647   0.35338919  0.213415025  0.628694804
## WoodDeckSF    -0.113414925   0.18970117  0.194809273  0.303309856
## OpenPorchSF   -0.112517508   0.28475555  0.221935283  0.441457956
## EnclosedPorch  0.017777155  -0.05083680 -0.065138425 -0.341520132
## X3SsnPorch    -0.026721950  -0.01547977  0.022247711  0.015156258
```

```
## ScreenPorch   -0.048762938   0.03662590  0.186523520 -0.086250685
## PoolArea      -0.015345274   0.05739577  0.095610078 -0.002183041
## MiscVal        0.014953729  -0.03538104 -0.040695422 -0.067630238
## MoSold         0.024545989   0.04394030  0.051081230  0.017966227
## YrSold         0.025878307  -0.03238604 -0.040779394  0.006837074
## SalePrice     -0.176453156   0.53840571  0.503902461  0.626219262
##                 GarageCars   GarageArea   WoodDeckSF  OpenPorchSF
## LotFrontage    0.341330327  0.370370104  0.102095760  0.179314052
## LotArea        0.341097718  0.355999054  0.168029666  0.189911247
## OverallQual    0.593647572  0.530278646  0.302275283  0.458852767
## OverallCond   -0.316159374 -0.260390407 -0.052770647 -0.200632164
## YearBuilt      0.601101226  0.538034142  0.299407161  0.442111039
## YearRemodAdd   0.492888530  0.439150316  0.261491485  0.384326792
## MasVnrArea     0.396932750  0.364521183  0.204879535  0.230627560
## BsmtFinSF1     0.133503072  0.200467636  0.174923779  0.086836515
## BsmtFinSF2    -0.110580365 -0.050814936  0.045529152 -0.090767878
## BsmtUnfSF      0.204513211  0.146350368 -0.021810468  0.163910596
## TotalBsmtSF    0.477102844  0.499281734  0.238384210  0.310282171
## X1stFlrSF      0.462872297  0.489323943  0.226807371  0.253944843
## X2ndFlrSF      0.139541434  0.081253975  0.090109383  0.200210428
## LowQualFinSF  -0.031851398 -0.014606569 -0.045574025  0.012486662
## GrLivArea      0.509524917  0.465393635  0.248919989  0.401168185
## BsmtFullBath   0.132195453  0.194232917  0.151001396  0.104174804
## BsmtHalfBath  -0.061671808 -0.049452152  0.062441372 -0.078639383
## FullBath       0.545065468  0.472821175  0.260847826  0.394750044
## HalfBath       0.188727249  0.114172960  0.112320643  0.246707280
## BedroomAbvGr   0.142951469  0.132319982  0.084445141  0.081954835
## KitchenAbvGr   0.053060375  0.037465647 -0.113414925 -0.112517508
## TotRmsAbvGrd   0.418079706  0.353389195  0.189701171  0.284755555
## Fireplaces     0.279547716  0.213415025  0.194809273  0.221935283
## GarageYrBlt    0.645409535  0.628694804  0.303309856  0.441457956
## GarageCars     1.000000000  0.838992979  0.260239830  0.365683339
## GarageArea     0.838992979  1.000000000  0.251880805  0.376673857
## WoodDeckSF     0.260239830  0.251880805  1.000000000  0.146594212
## OpenPorchSF    0.365683339  0.376673857  0.146594212  1.000000000
## EnclosedPorch -0.215437648 -0.182923212 -0.142213773 -0.208606090
## X3SsnPorch     0.017798899  0.017395654 -0.044511050  0.008117841
## ScreenPorch    0.004322419  0.008600182 -0.099112821  0.014074926
## PoolArea       0.014738770  0.052321922  0.039950711  0.023601592
## MiscVal       -0.070198330 -0.032377680 -0.002552037 -0.034819899
## MoSold         0.055704107  0.041401128  0.061306262  0.074300709
## YrSold        -0.032516074 -0.010131796  0.015724947 -0.051979894
## SalePrice      0.668072170  0.631302788  0.366395457  0.507618069
##                EnclosedPorch   X3SsnPorch  ScreenPorch     PoolArea
## LotFrontage    -0.097629247  0.062389872  0.036719900  0.086092162
## LotArea        -0.079932870  0.044701006  0.106828879  0.086850564
## OverallQual    -0.200523384  0.015475303  0.032929482  0.065780007
## OverallCond     0.110328966  0.009856910  0.107517740 -0.023172667
## YearBuilt      -0.415353875  0.019366788 -0.079994041  0.010569106
## YearRemodAdd   -0.267526111  0.030931795 -0.038799487  0.009272607
```

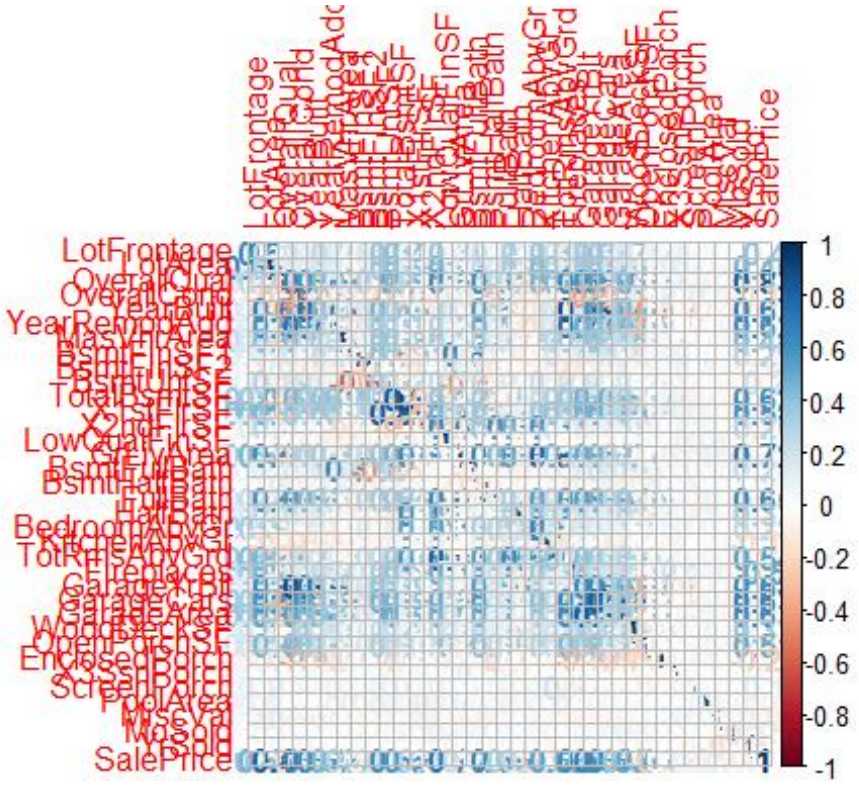
```
## MasVnrArea     -0.196440656  0.044434329  0.033942816  0.014386412
## BsmtFinSF1     -0.151555666  0.036169664  0.078383880  0.086986654
## BsmtFinSF2      0.037317210 -0.024518933  0.039760171  0.091224798
## BsmtUnfSF       0.018886754  0.022902517 -0.014195470 -0.056655367
## TotalBsmtSF    -0.195629093  0.048304585  0.079378295  0.070588661
## X1stFlrSF      -0.156584494  0.050413909  0.101152573  0.070350280
## X2ndFlrSF       0.057740311 -0.026526178  0.017292097  0.066411470
## LowQualFinSF    0.016679304  0.041134161 -0.009029495  0.091201766
## GrLivArea      -0.071554944  0.016367017  0.089095467  0.065229280
## BsmtFullBath   -0.069209765 -0.004963405  0.030869612  0.097319847
## BsmtHalfBath   -0.036650395  0.094022292  0.009361298  0.036107514
## FullBath       -0.196792395  0.036521668 -0.026848815  0.040220228
## HalfBath       -0.117676999  0.002922579  0.058725619  0.043386207
## BedroomAbvGr    0.006239924 -0.019547495  0.051641963  0.070511913
## KitchenAbvGr    0.017777155 -0.026721950 -0.048762938 -0.015345274
## TotRmsAbvGrd   -0.050836796 -0.015479774  0.036625903  0.057395772
## Fireplaces     -0.065138425  0.022247711  0.186523520  0.095610078
## GarageYrBlt    -0.341520132  0.015156258 -0.086250685 -0.002183041
## GarageCars     -0.215437648  0.017798899  0.004322419  0.014738770
## GarageArea     -0.182923212  0.017395654  0.008600182  0.052321922
## WoodDeckSF     -0.142213773 -0.044511050 -0.099112821  0.039950711
## OpenPorchSF    -0.208606090  0.008117841  0.014074926  0.023601592
## EnclosedPorch   1.000000000 -0.033380271 -0.073900666  0.010196863
## X3SsnPorch     -0.033380271  1.000000000 -0.038819591 -0.009370566
## ScreenPorch    -0.073900666 -0.038819591  1.000000000  0.025105110
## PoolArea        0.010196863 -0.009370566  0.025105110  1.000000000
## MiscVal         0.049566493  0.019789381  0.044549580  0.060087423
## MoSold         -0.060959277  0.028781407  0.013141836 -0.040743258
## YrSold          0.011727778  0.004975673  0.008075427 -0.053332198
## SalePrice      -0.241361901  0.044933771  0.100215658  0.052573200
##                    MiscVal        MoSold       YrSold   SalePrice
## LotFrontage    0.020739437  0.0305732569  0.007657252  0.39965695
## LotArea        0.047015162  0.0117359217 -0.025307127  0.46028741
## OverallQual   -0.107329952  0.0799273853 -0.010309413  0.82710824
## OverallCond    0.108376809 -0.0155413220  0.041029181 -0.17495493
## YearBuilt     -0.117302815  0.0213327029 -0.004651017  0.67019837
## YearRemodAdd  -0.100067923  0.0256763370  0.050791355  0.61731119
## MasVnrArea    -0.094631831  0.0290485037 -0.011500579  0.43544368
## BsmtFinSF1     0.010187895 -0.0089253379  0.014983327  0.28532979
## BsmtFinSF2     0.018661546 -0.0602866436  0.036695437 -0.07609782
## BsmtUnfSF     -0.042736789  0.0254578108 -0.027446064  0.20158315
## TotalBsmtSF   -0.063806227  0.0194265979 -0.008978755  0.61829789
## X1stFlrSF     -0.047238000  0.0449918826 -0.015761290  0.58292981
## X2ndFlrSF     -0.029978373  0.0514682375 -0.021767207  0.26369234
## LowQualFinSF   0.062295561 -0.0006644359 -0.019216380 -0.05341963
## GrLivArea     -0.079368880  0.0851533307 -0.016019501  0.72497669
## BsmtFullBath   0.016346189 -0.0312064611  0.055227663  0.22800120
## BsmtHalfBath   0.003149930  0.0332711266 -0.047505384 -0.03940044
## FullBath      -0.083238890  0.0799841890  0.008607973  0.65115792
## HalfBath      -0.062615754  0.0027012937 -0.018867453  0.31991762
```

```
## BedroomAbvGr   -0.008164026  0.0346319446 -0.026042899  0.23498624
## KitchenAbvGr    0.014953729  0.0245459894  0.025878307 -0.17645316
## TotRmsAbvGrd   -0.035381040  0.0439403045 -0.032386036  0.53840571
## Fireplaces     -0.040695422  0.0510812301 -0.040779394  0.50390246
## GarageYrBlt    -0.067630238  0.0179662273  0.006837074  0.62621926
## GarageCars     -0.070198330  0.0557041072 -0.032516074  0.66807217
## GarageArea     -0.032377680  0.0414011276 -0.010131796  0.63130279
## WoodDeckSF     -0.002552037  0.0613062616  0.015724947  0.36639546
## OpenPorchSF    -0.034819899  0.0743007085 -0.051979894  0.50761807
## EnclosedPorch   0.049566493 -0.0609592771  0.011727778 -0.24136190
## X3SsnPorch      0.019789381  0.0287814070  0.004975673  0.04493377
## ScreenPorch     0.044549580  0.0131418364  0.008075427  0.10021566
## PoolArea        0.060087423 -0.0407432575 -0.053332198  0.05257320
## MiscVal         1.000000000  0.0156824735  0.036579022 -0.09190581
## MoSold          0.015682473  1.0000000000 -0.154791117  0.08191427
## YrSold          0.036579022 -0.1547911173  1.000000000 -0.01765020
## SalePrice      -0.091905807  0.0819142656 -0.017650201  1.00000000
```

```
corrplot(cor(housedata_num,use="complete.obs", method='spearman'),
method="ellipse", cex.axis=0.8)
```




```
corrplot(cor(housedata_num,use="complete.obs", method='spearman'),
method="number", cex.axis=0.8)
```


LotFrontage
SalePrice
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1