Lab No. 6

Elizabeth Walter

```
knitr::opts_chunk$set(echo = TRUE,
                      fig.align = 'center')
```

```
library(tidyverse)  # For ggplot, dplyr, and friends
library(patchwork)  # For combining ggplot plots
library(GGally)     # For scatterplot matrices
library(broom)      # For converting model objects to data frames

ameslist <- read.csv("C:/Users/walte/Desktop/MSU SSQDA/SSC 442/Data/ames.csv",
                 header = TRUE,
                 sep = ",")
```

1. Prune the data to all of the variables that are type = int about which you have some reasonable intuition for what they mean. This must include the variable SalePrice. Save this new dataset as Ames. Produce documentation for this object in the form of a .txt file. This must describe each of the preserved variables, the values it can take (e.g., can it be negative?) and your interpretation of the variable.

```
# list class of each var in ameslist & get indexes of vars in ameslist where class = int
b <- which(sapply(ameslist, class) %in% c('integer'))
length(b) # expect 38 variables
```

```
## [1] 38
```

```
# get list of vars in ameslist of type int
names(ameslist[b])
```

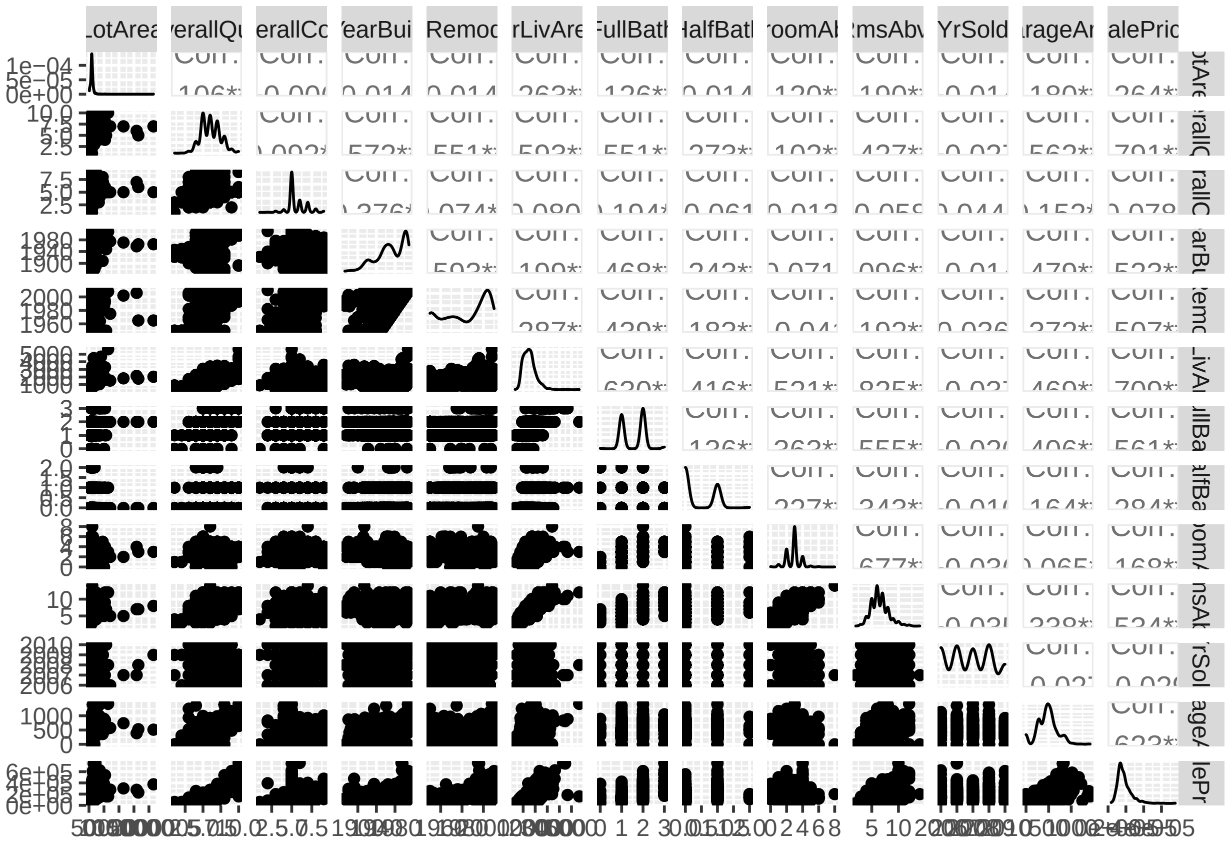
```
##  [1] "Id"            "MSSubClass"    "LotFrontage"   "LotArea"
##  [5] "OverallQual"   "OverallCond"   "YearBuilt"     "YearRemodAdd"
##  [9] "MasVnrArea"    "BsmtFinSF1"    "BsmtFinSF2"    "BsmtUnfSF"
## [13] "TotalBsmtSF"   "X1stFlrSF"     "X2ndFlrSF"     "LowQualFinSF"
## [17] "GrLivArea" 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

```r
# get new dataframe of only variables of type int
Ames <- ameslist[names(ameslist[b])]

q1 = read.csv("C:/Users/walte/Desktop/MSU SSQDA/SSC 442/Labs/Lab 6/Q1 documentation",
              header=TRUE, sep = ",")
q1
```

```
##    Variable.Name Description                    Interpretation
## 1             Id     Nominal            House identification #
## 2     MSSubClass     Nominal         House Classification Code
## 3    LotFrontage  Continuous   Length street touching property
## 4        LotArea  Continuous                          Lot size
## 5    OverallQual     Ordinal   Rating overall quality of house
## 6    OverallCond     Ordinal Rating overall condition of house
## 7      YearBuilt    Discrete           Year property was built
## 8   YearRemodAdd    Discrete        year remodel/addition done
## 9     MasVnrArea  Continuous                        area of ??
## 10    BsmtFinSF1  Continuous       Finished area of bsmt(ft^2)
## 11    BsmtFinSF2  Continuous Diff finished area of bsmt?(ft^2)
## 12     BsmtUnfSF  Continuous     Unfinished area of bsmt(ft^2)
## 13   TotalBsmtSF  Continuous      Total area of basement(ft^2)
## 14     X1stFlrSF  Continuous         Area of first floor(ft^2)
## 15     X2ndFlrSF  Continuous        Area of second floor(ft^2)
## 16  LowQualFinSF  Continuous         Low Quality ?? Area(ft^2)
## 17     GrLivArea  Continuous    Combined area 1st & 2nd floors
## 18  BsmtFullBath    Discrete      # Full bathrooms in basement
## 19  BsmtHalfBath    Discrete      # Half bathrooms in basement
## 20      FullBath    Discrete       # Full baths above basement
## 21      HalfBath    Discrete       # Half baths above basement
## 22  BedroomAbvGr    Discrete         # bedrooms above basement
## 23  KitchenAbvGr    Discrete         # Kitchens above basement
## 24  TotRmsAbvGrd    Discrete      Total # rooms above basement
## 25    Fireplaces    Discrete                      # Fireplaces
## 26   GarageYrBlt    Discrete                 Year Garage Built
## 27    GarageCars    Discrete              # cars fit in garage
## 28    GarageArea  Continuous                    Area of garage
## 29    WoodDeckSF  Continuous           Area of wood deck(ft^2)
## 30   OpenPorchSF  Continuous         Area of open porch?(ft^2)
## 31 EnclosedPorch  Continuous           Area of enclosed porch?
## 32    X3SsnPorch  Continuous     Area of another type of porch
## 33   ScreenPorch  Continuous         Area of screened-in porch
## 34      PoolArea  Continuous                      Area of Pool
## 35       MiscVal  Continuous                Value of something
## 36        MoSold    Discrete                        Month Sold
## 37        YrSold    Discrete                         Year Sold
## 38     SalePrice  Continuous                 Price sold for($)
##          Values.Can.Take
## 1        int from 1-1460
## 2  mult of 5 btwn 20-190
## 3   0 - all pos integers
## 4           all pos ints
## 5                 1 - 10
## 6                 1 - 10
## 7               any year
## 8               any year
## 9       0 - all pos ints
## 10      0 - all pos ints
## 11      0 - all pos ints
## 12      0 - all pos ints
## 13      0 - all pos ints
## 14      0 - all pos ints
## 15      0 - all pos ints
## 16      0 - all pos ints
## 17      0 - all pos ints
## 18      0 - all pos ints
## 19      0 - all pos ints
## 20      0 - all pos ints
## 21      0 - all pos ints
## 22      0 - all pos ints
## 23      0 - all pos ints
## 24          all pos ints
## 25      0 - all pos ints
## 26              any year
## 27      0 - all pos ints
## 28      0 - all pos ints
## 29      0 - all pos ints
## 30      0 - all pos ints
## 31      0 - all pos ints
## 32      0 - all pos ints
## 33      0 - all pos ints
## 34      0 - all pos ints
## 35      0 - all pos ints
## 36                1 - 12
## 37   Any year>year built
## 38      0 - all pos ints
```

2. Produce a scatterplot matrix which includes 12 of the variables that are type = int in the data set. Choose those that you believe are likely to be correlated with SalePrice.

```
price_corrs <- Ames %>%
  select(LotArea, OverallQual, OverallCond, YearBuilt, YearRemodAdd, GrLivArea,
         FullBath, HalfBath, BedroomAbvGr, TotRmsAbvGrd, YrSold, GarageArea, SalePrice)

ggpairs(price_corrs)
```



3. Compute a matrix of correlations between these variables using the function cor(). Does this match your prior beliefs? Briefly discuss the correlation between the miscellaneous variables and SalePrice.

All of the variables that are positively correlated with SalePrice match my prior beliefs of what their direction of correlation would be. OverallQual has the strongest correlation to SalePrice (both negative and positive) of all the variables chosen, with a positive correlation of r = 0.79. I believed that there would be a high positive correlation between those two variables. GrLivArea also has a strong positive correlation (r = 0.71), which I expected. I thought LotArea and SalePrice would have a stronger positive correlation than r = 0.26. GarageArea and SalePrice have a stronger correlation than I expected (r = 0.62), but I expected it to be moderate and positive. I am surprised that BedroomAbvGr and SalePrice are not strongly correlated (r = 0.17), and more so that it is weaker than the correlation between HalfBath and SalePrice (r = 0.28). I would have expected it to be similar to, if not greater than, the correlation between SalePrice and FullBath, which I also expected to be a strong correlation, but is slightly weaker than expected (r = 0.56). The correlation between YearBuilt and SalePrice is similar to what I expected (r = 0.52).

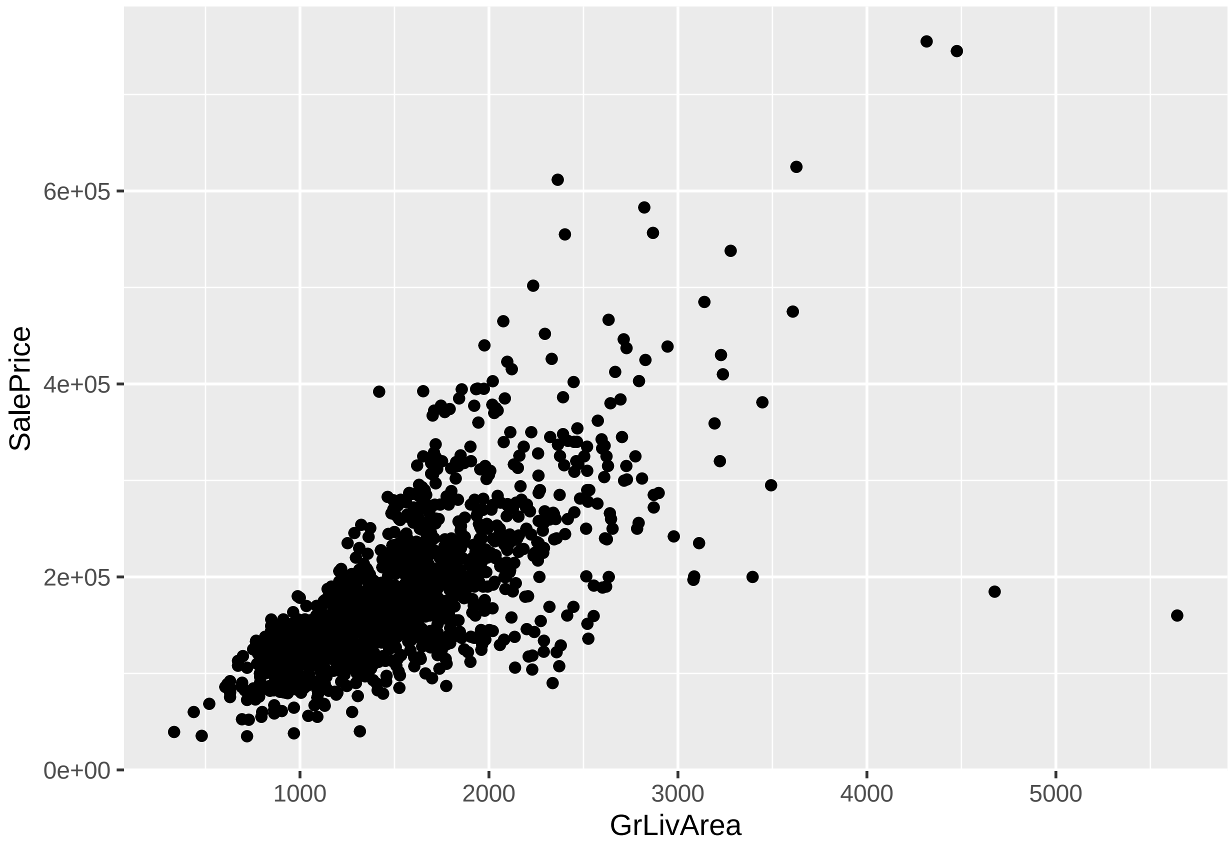
The only negatively correlated variables are also pretty weakly correlated. I thought that YrSold might be more correlated with SalePrice, but it appears to have no correlation (r = - 0.03), and I expected OverallCond and SalePrice to have a positive and stronger correlation than the very weak, negative correlation they have (r = -0.08).

```
ames_corr <- Ames %>%
  select(LotArea, OverallQual, OverallCond, YearBuilt, YearRemodAdd, GrLivArea,
         FullBath, HalfBath, BedroomAbvGr, TotRmsAbvGrd, YrSold, GarageArea, SalePrice) %>%
  cor()

ames_corr
```

```
##                  LotArea OverallQual OverallCond   YearBuilt YearRemodAdd
## LotArea       1.00000000  0.10580574 -0.00563627  0.01422765   0.01378843
## OverallQual   0.10580574  1.00000000 -0.09193234  0.57232277   0.55068392
## OverallCond  -0.00563627 -0.09193234  1.00000000 -0.37598320   0.07374150
## YearBuilt     0.01422765  0.57232277 -0.37598320  1.00000000   0.59285498
## YearRemodAdd  0.01378843  0.55068392  0.07374150  0.59285498   1.00000000
## GrLivArea     0.26311617  0.59300743 -0.07968587  0.19900971   0.28738852
## FullBath      0.12603063  0.55059971 -0.19414949  0.46827079   0.43904648
## HalfBath      0.01425947  0.27345810 -0.06076933  0.24265591   0.18333061
## BedroomAbvGr  0.11968991  0.10167636  0.01298006 -0.07065122  -0.04058093
## TotRmsAbvGrd  0.19001478  0.42745234 -0.05758317  0.09558913   0.19173982
## YrSold       -0.01426141 -0.02734671  0.04394975 -0.01361768   0.03574325
## GarageArea    0.18040276  0.56202176 -0.15152137  0.47895382   0.37159981
## SalePrice     0.26384335  0.79098160 -0.07785589  0.52289733   0.50710097
##                GrLivArea    FullBath    HalfBath BedroomAbvGr TotRmsAbvGrd
## LotArea       0.26311617  0.12603063  0.01425947   0.11968991   0.19001478
## OverallQual   0.59300743  0.55059971  0.27345810   0.10167636   0.42745234
## OverallCond  -0.07968587 -0.19414949 -0.06076933   0.01298006  -0.05758317
## YearBuilt     0.19900971  0.46827079  0.24265591  -0.07065122   0.09558913
## YearRemodAdd  0.28738852  0.43904648  0.18333061  -0.04058093   0.19173982
## GrLivArea     1.00000000  0.63001165  0.41577164   0.52126951   0.82548937
## FullBath      0.63001165  1.00000000  0.13638059   0.36325198   0.55478425
## HalfBath      0.41577164  0.13638059  1.00000000   0.22665148   0.34341486
## BedroomAbvGr  0.52126951  0.36325198  0.22665148   1.00000000   0.67661994
## TotRmsAbvGrd  0.82548937  0.55478425  0.34341486   0.67661994   1.00000000
## YrSold       -0.03652582 -0.01966884 -0.01026867  -0.03601389  -0.03451635
## GarageArea    0.46899748  0.40565621  0.16354936   0.06525253   0.33782212
## SalePrice     0.70862448  0.56066376  0.28410768   0.16821315   0.53372316
##                   YrSold  GarageArea   SalePrice
## LotArea      -0.01426141  0.18040276  0.26384335
## OverallQual  -0.02734671  0.56202176  0.79098160
## OverallCond   0.04394975 -0.15152137 -0.07785589
## YearBuilt    -0.01361768  0.47895382  0.52289733
## YearRemodAdd  0.03574325  0.37159981  0.50710097
## GrLivArea    -0.03652582  0.46899748  0.70862448
## FullBath     -0.01966884  0.40565621  0.56066376
## HalfBath     -0.01026867  0.16354936  0.28410768
## BedroomAbvGr -0.03601389  0.06525253  0.16821315
## TotRmsAbvGrd -0.03451635  0.33782212  0.53372316
## YrSold        1.00000000 -0.02737794 -0.02892259
## GarageArea   -0.02737794  1.00000000  0.62343144
## SalePrice    -0.02892259  0.62343144  1.00000000
```

4. Produce a scatterplot between SalePrice and GrLivArea. Run a linear model using lm() to explore the relationship. Finally, use the abline() function to plot the relationship that you've found in the simple linear regression.

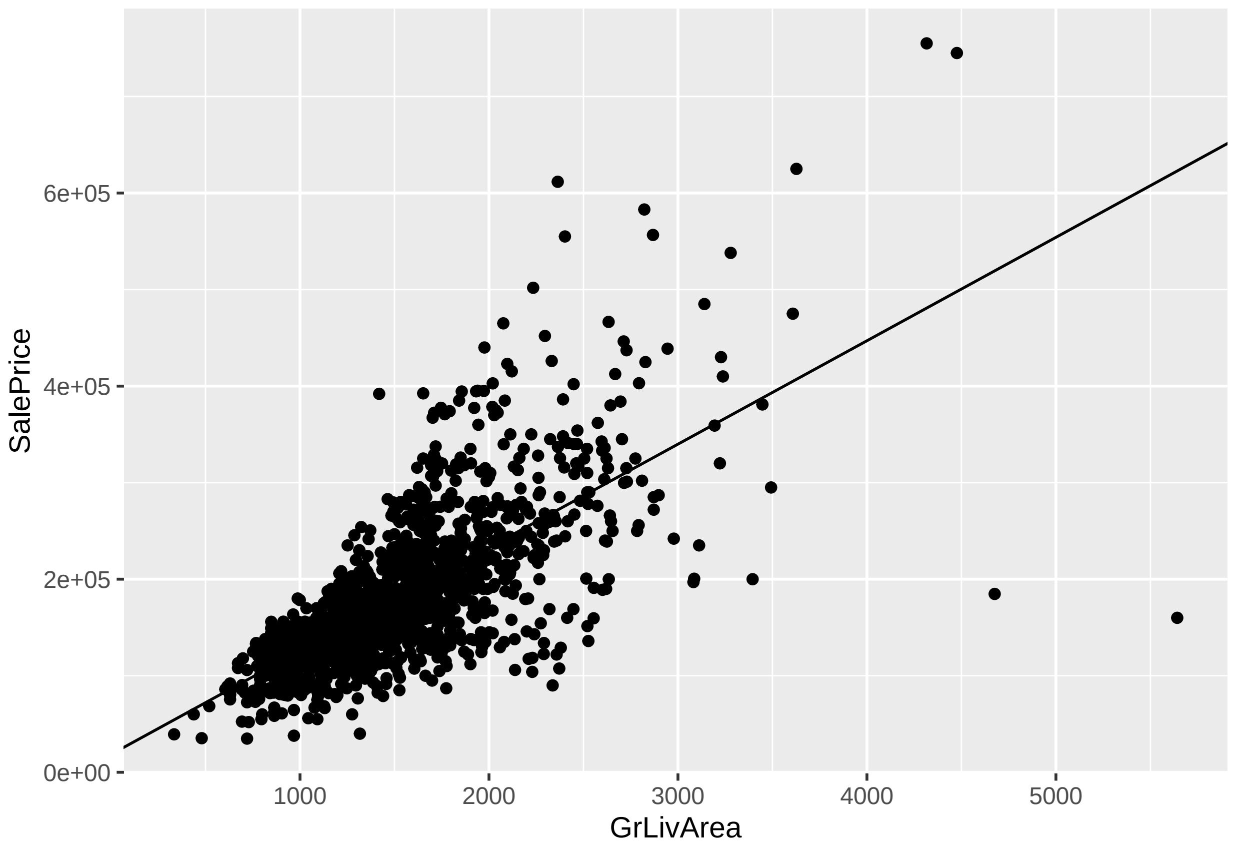
```
ggplot(Ames, aes(x = GrLivArea, y = SalePrice)) +
  geom_point()
```



```
price_livarea_reg <- lm(SalePrice ~ GrLivArea, data = Ames)
tidy(price_livarea_reg, conf.int = TRUE)
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic   p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 (Intercept)   18569.   4481.        4.14 3.61e-  5    9780.    27358.
## 2 GrLivArea       107.      2.79     38.3  4.52e-223     102.      113.
```

```
ggplot(Ames, aes(x = GrLivArea, y = SalePrice)) +
  geom_point() +
  geom_abline(intercept = 18569.0, slope = 107.1)
```



4b. What is the largest outlier that is above the regression line? Produce the other information about this house.

Largest outlier ABOVE the regression line is a house with GrLivArea of 4316 and SalePrice of $755,000

```
out_above <- Ames[which.max(Ames$SalePrice),]
out_above
```

```
##      Id MSSubClass LotFrontage LotArea OverallQual OverallCond YearBuilt
## 692 692         60         104   21535          10           6      1994
##     YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
## 692         1995       1170       1455          0       989        2444
##     X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath
## 692      2444      1872            0      4316            0            1
##     FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces
## 692        3        1            4            1           10          2
##     GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
## 692        1994          3        832        382          50             0
##     X3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold SalePrice
## 692          0           0        0       0      1   2007    755000
```

Largest outlier BELOW the regression line (also house with greatest distance from regression line) is a house with GrLivArea of 5642 and SalePrice of $160,000.

```
out_below <- Ames[which.max(Ames$GrLivArea),]
out_below
```

```
##        Id MSSubClass LotFrontage LotArea OverallQual OverallCond YearBuilt
## 1299 1299         60         313   63887          10           5      2008
##      YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
## 1299         2008        796       5644          0       466        6110
##      X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath
## 1299      4692       950            0      5642            2            0
##      FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces
## 1299        2        1            3            1           12          3
##      GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
## 1299        2008          2       1418        214         292             0
##      X3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold SalePrice
## 1299          0           0      480       0      1   2008    160000
```