house-prices.R

sharmaam

Mon Oct 3 15:43:17 2016

```
##############################################################################
# Learning Statistical Data Analysis using                                   #
# Kaggle data for House Prices                                               #
# https://www.kaggle.com/c/house-prices-advanced-regression-techniques       #
# author: Amit Sharma                                                        #
##############################################################################

rm(list=ls())

# Set the working directory
setwd('~/axs/kaggle/house-prices/')

# Let's set scientific notation
options(scipen=10)

# Load the dataframe from the CSV file
data = read.csv('data/train.csv')

# Initialize a few color pallettes
myCols = rainbow(8:13)

##################################################
#           EXPLORATORY ANALYSIS                 #
##################################################

# Column types
str(data)
```

```
## 'data.frame':    1460 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : 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 ...
```

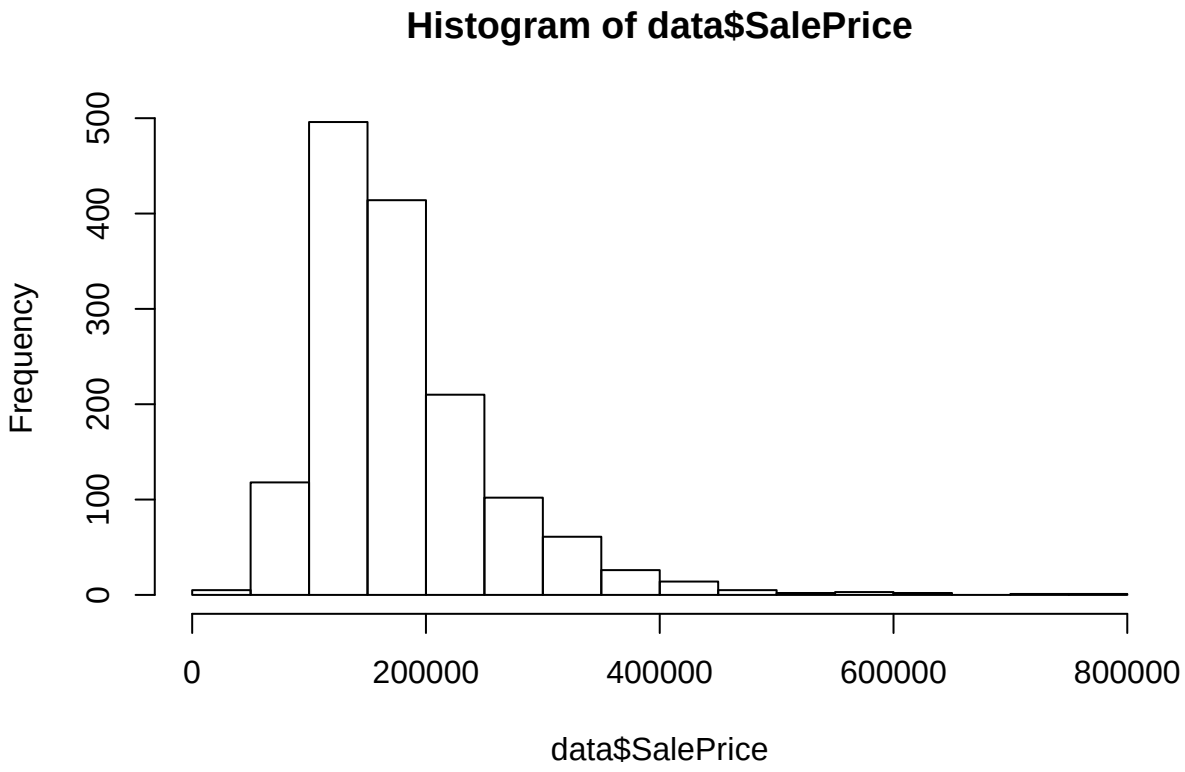
```r
# Column Summaries
summary(data)
```

```
##        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
```

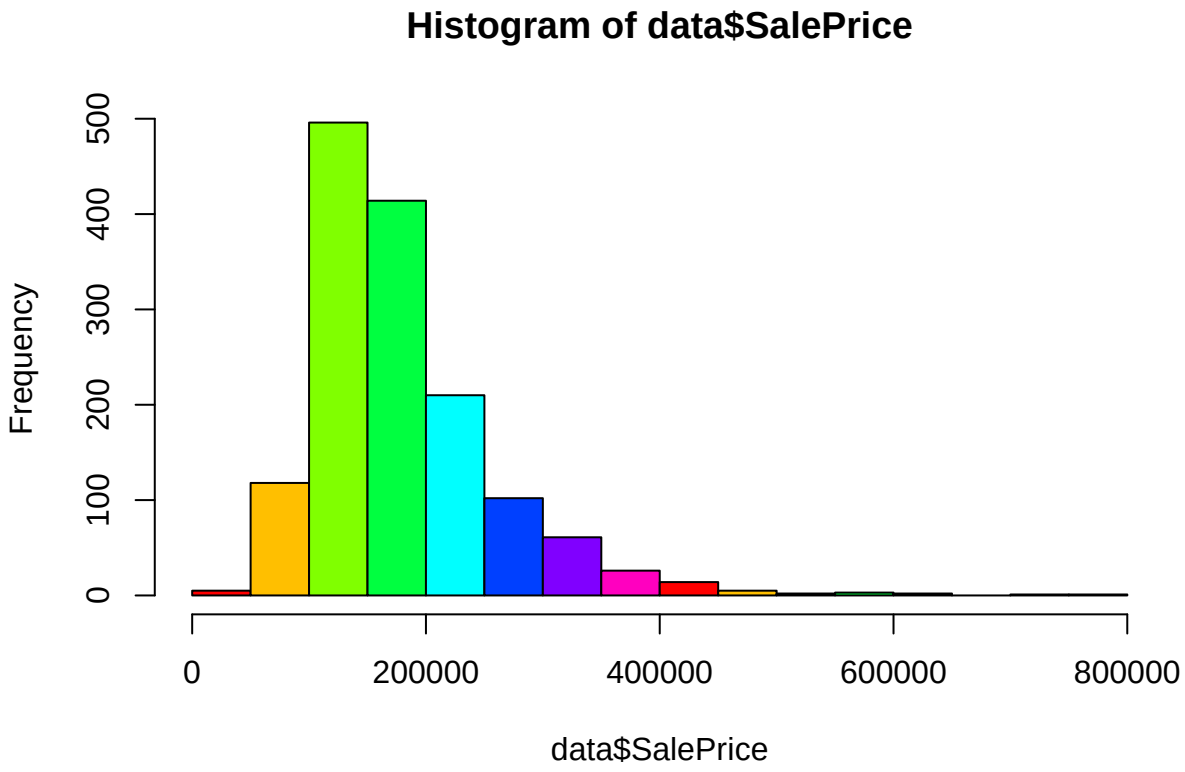
```
##  CompShg:1434   VinylSd:515   VinylSd:504   BrkCmn : 15   Min.   :   0.0
##  Tar&Grv:  11   HdBoard:222   MetalSd:214   BrkFace:445   1st Qu.:   0.0
##  WdShngl:   6   MetalSd:220   HdBoard:207   None   :864   Median :   0.0
##  WdShake:   5   Wd Sdng:206   Wd Sdng:197   Stone  :128   Mean   : 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
```

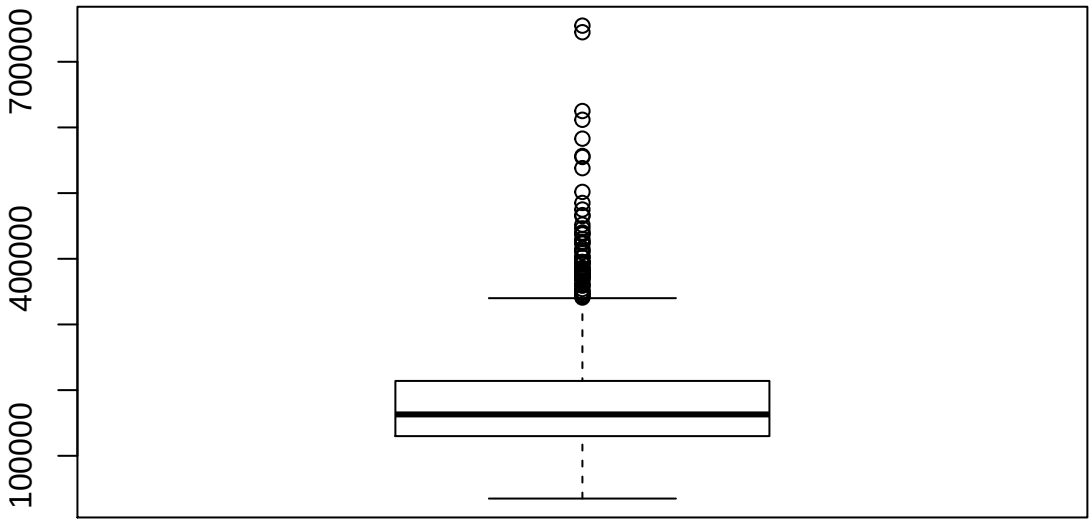
```r
# Let's check the price distribution using a histogram
hist(data$SalePrice)
```



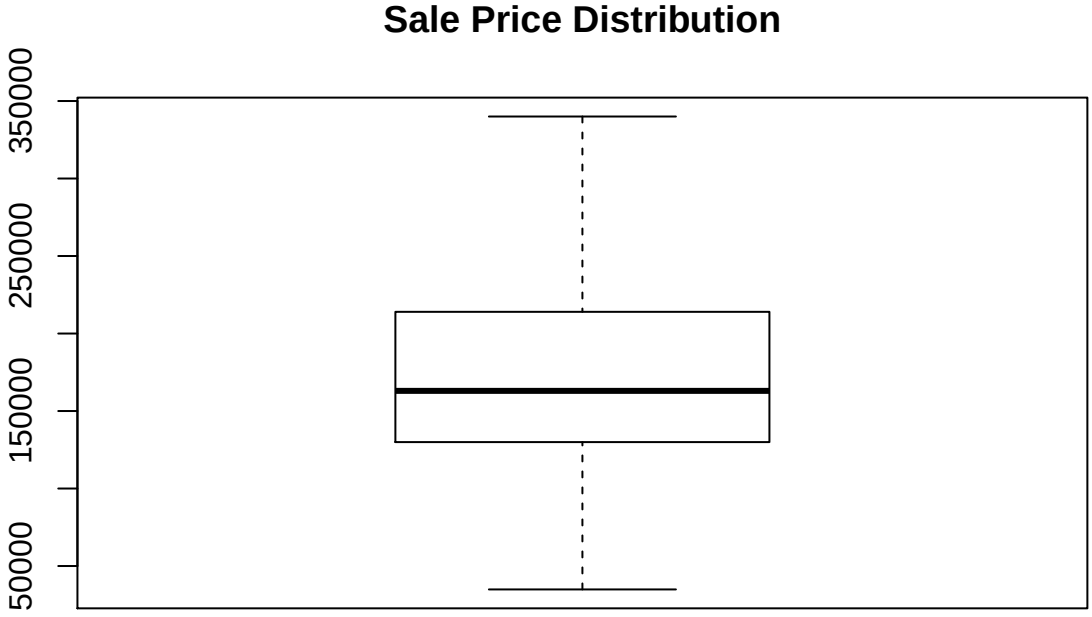

```
# Let's color it up a bit
hist(data$SalePrice, col=myCols)
```



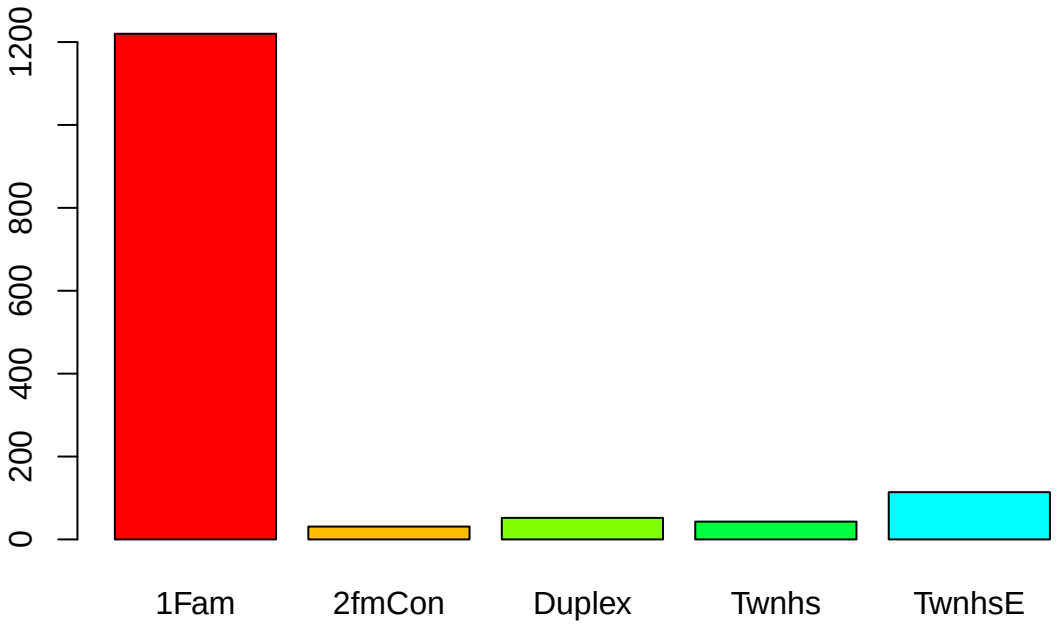
```r
# Another way to explore the price distribution
boxplot(data$SalePrice)
```



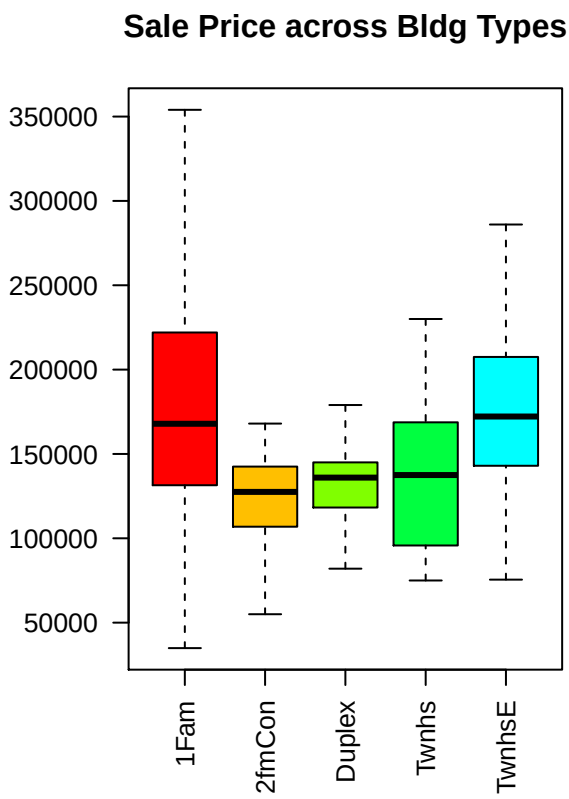

```r
# Get rid of outliers on the boxplot
boxplot(data$SalePrice, outline=F, main="Sale Price Distribution")
```

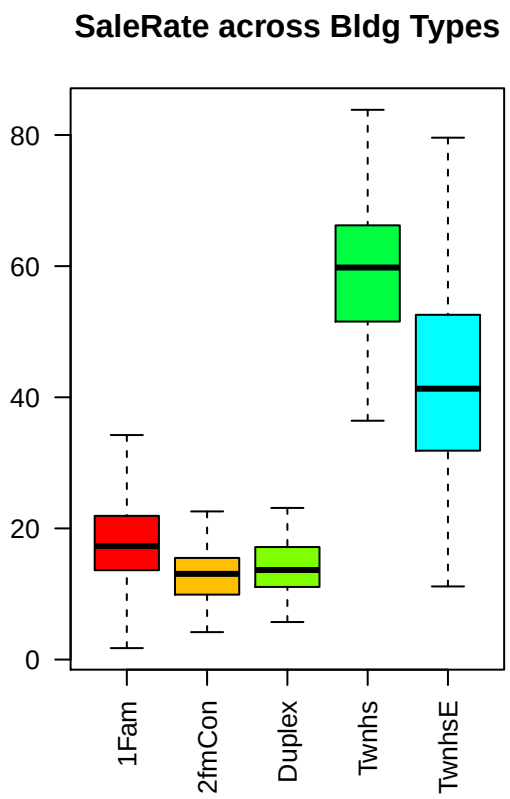



```r
# If we are dealing with just one data frame, it helps to attach it so
# we can just use the column names
attach(data)

# Lets see how many Bldg Types are there
barplot(table(BldgType), col=myCols)
```

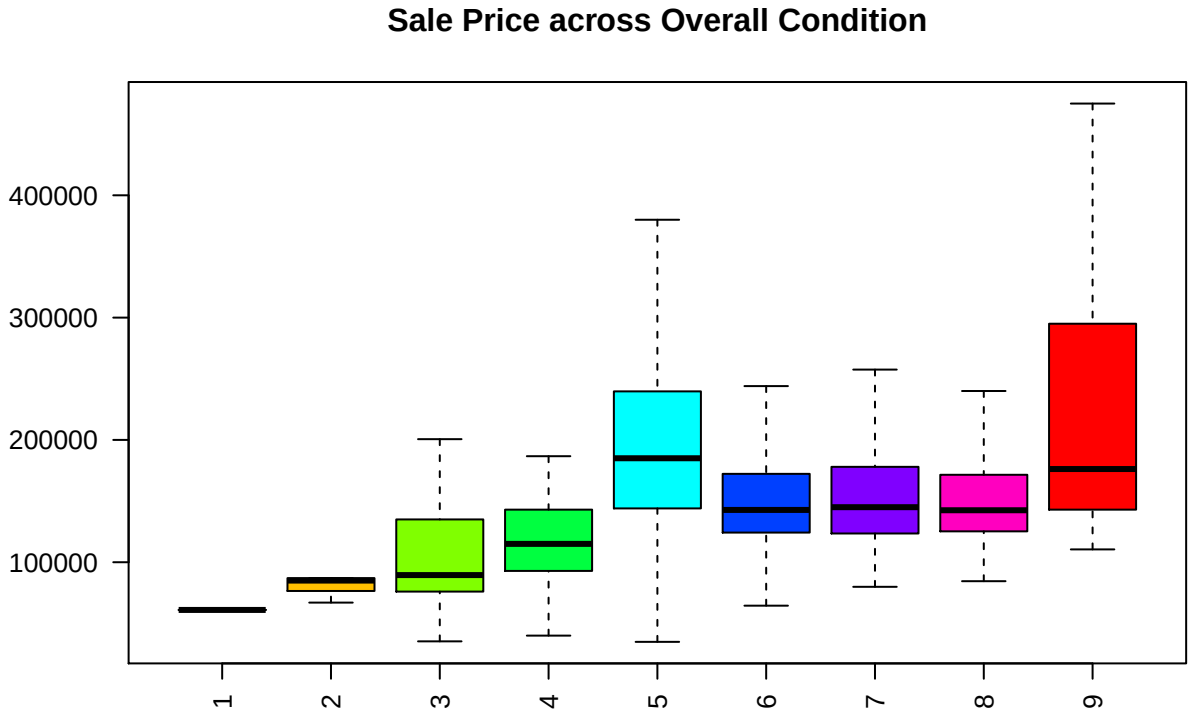



```
par(mfrow=c(1,2), cex=0.8, las=2)
# Lets check the prices boxplot across Bldg Types
boxplot(SalePrice ~ BldgType, col=myCols, outline=F,
        main="Sale Price across Bldg Types")

# Lets craete a new column for House Rate
SaleRate = SalePrice / LotArea

# Lets check the prices boxplot across Bldg Types
boxplot(SaleRate ~ BldgType, col=myCols,
        outline=F, main="SaleRate across Bldg Types")
```





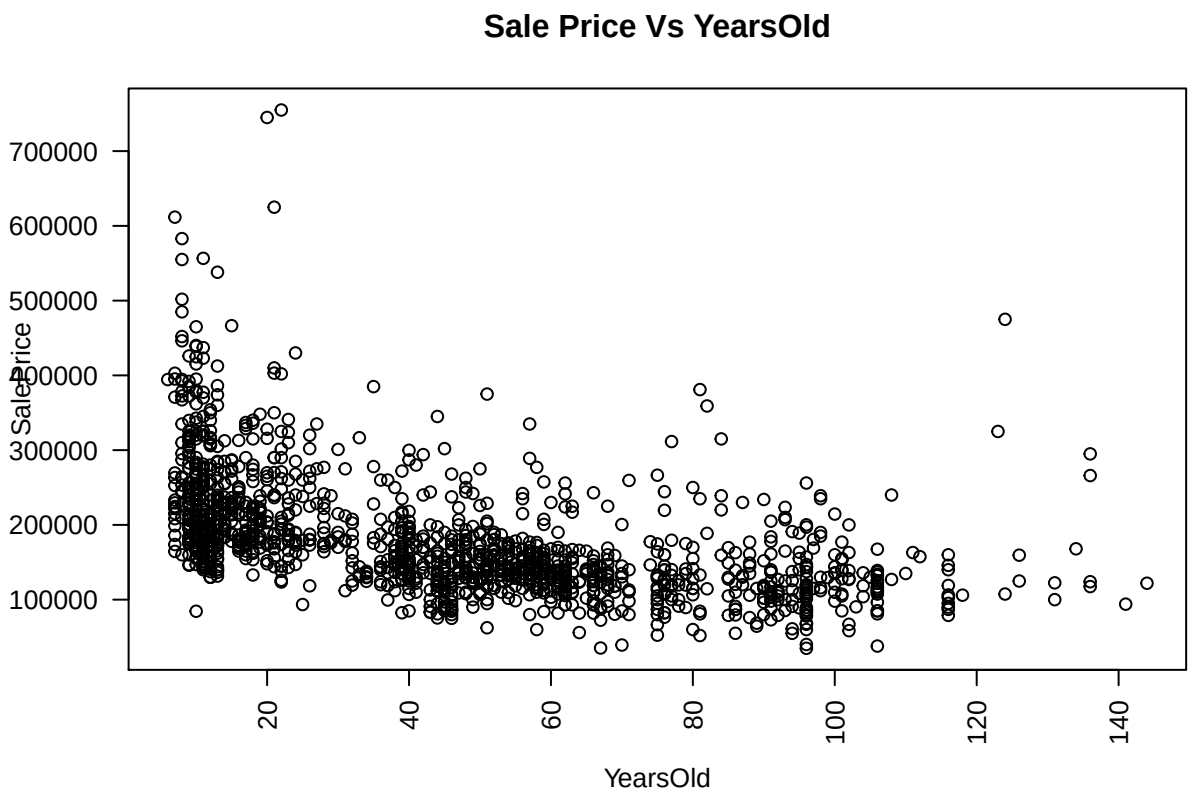

```
par(mfrow=c(1,1), cex=0.8, las=2)
boxplot(SalePrice ~ OverallCond, col=myCols, outline=F,
        main="Sale Price across Overall Condition")
```



```r
# Calculate Years Old
data$YearsOld = 2016 - YearBuilt
attach(data)
```

```
## The following objects are masked from data (pos = 3):
##
##     Alley, BedroomAbvGr, BldgType, BsmtCond, BsmtExposure,
##     BsmtFinSF1, BsmtFinSF2, BsmtFinType1, BsmtFinType2,
##     BsmtFullBath, BsmtHalfBath, BsmtQual, BsmtUnfSF, CentralAir,
##     Condition1, Condition2, Electrical, EnclosedPorch, ExterCond,
##     Exterior1st, Exterior2nd, ExterQual, Fence, FireplaceQu,
##     Fireplaces, Foundation, FullBath, Functional, GarageArea,
##     GarageCars, GarageCond, GarageFinish, GarageQual, GarageType,
##     GarageYrBlt, GrLivArea, HalfBath, Heating, HeatingQC,
##     HouseStyle, Id, KitchenAbvGr, KitchenQual, LandContour,
##     LandSlope, LotArea, LotConfig, LotFrontage, LotShape,
##     LowQualFinSF, MasVnrArea, MasVnrType, MiscFeature, MiscVal,
##     MoSold, MSSubClass, MSZoning, Neighborhood, OpenPorchSF,
##     OverallCond, OverallQual, PavedDrive, PoolArea, PoolQC,
##     RoofMatl, RoofStyle, SaleCondition, SalePrice, SaleType,
##     ScreenPorch, Street, TotalBsmtSF, TotRmsAbvGrd, Utilities,
##     WoodDeckSF, X1stFlrSF, X2ndFlrSF, X3SsnPorch, YearBuilt,
##     YearRemodAdd, YrSold
```

```r
plot (YearsOld, SalePrice, main="Sale Price Vs YearsOld")
```



```r
# Lets just look at Numeric columns for a while for
# correlation analysis
numericCols = sapply(data,is.numeric)
data = data[,numericCols]

# Check correlation of marks
# cr = round(cor(data), 1)
#
# # Using library corrplot
# library(corrplot)
# corrplot(cr, type="lower")

# Lets remove the NAs
data = data[complete.cases(data),]

# Combining plots
#par(mfrow=c(1,2), cex=0.8, las=2)
#boxplot(marksDF, outline=F, col=rainbow(12:14))
#barplot(sapply(marksDF, mean), col=rainbow(11:14))

##########################
# REGRESSSION ANALYSIS  #
##########################

# Lets take some students as test set and remaining as test
N = nrow(data)
# Randomly sample 90% of records to be used for train
trainRows = sample(1:N, N*0.9, replace=F)
# Use the sample to separate train/test from df
train = data[trainRows,]
test = data[-trainRows,]

# Now lets fit the model
model = lm(SalePrice~.,
           data=train)
summary(model)
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -428565  -17599   -2529   15519  318771
##
## Coefficients: (3 not defined because of singularities)
##                   Estimate   Std. Error t value         Pr(>|t|)
## (Intercept)   -570151.2545 1828094.3765  -0.312         0.755196
## Id                 -1.1285       2.8599  -0.395         0.693225
## MSSubClass       -193.5151      36.6658  -5.278 0.00000016124646 ***
## LotFrontage      -116.9546      65.2545  -1.792         0.073398 .
## LotArea             0.5431       0.1615   3.362         0.000804 ***
## OverallQual     19392.8777    1602.9213  12.098          < 2e-16 ***
```

```
## OverallCond      4884.8706    1470.1969   3.323         0.000925 ***
## YearBuilt         303.4677      93.5482   3.244         0.001219 **
## YearRemodAdd       90.1310      92.1122   0.978         0.328075
## MasVnrArea         31.7461       7.4801   4.244 0.00002404869796 ***
## BsmtFinSF1         14.1752       6.1874   2.291         0.022178 *
## BsmtFinSF2          7.3623       9.2755   0.794         0.427543
## BsmtUnfSF           3.4979       5.6308   0.621         0.534606
## TotalBsmtSF             NA           NA      NA               NA
## X1stFlrSF          47.9899       7.8130   6.142 0.00000000118427 ***
## X2ndFlrSF          45.6749       6.4958   7.031 0.00000000000385 ***
## LowQualFinSF       55.6876      30.7076   1.813         0.070066 .
## GrLivArea               NA           NA      NA               NA
## BsmtFullBath     8932.1311    3399.4526   2.628         0.008736 **
## BsmtHalfBath     2688.4673    5556.5427   0.484         0.628610
## FullBath         5714.2470    3763.0274   1.519         0.129208
## HalfBath          451.3421    3582.1229   0.126         0.899759
## BedroomAbvGr    -9851.4600    2319.1449  -4.248 0.00002365333873 ***
## KitchenAbvGr   -21479.0530    7285.5993  -2.948         0.003273 **
## TotRmsAbvGrd     5260.2219    1597.7274   3.292         0.001029 **
## Fireplaces       4549.1454    2343.3003   1.941         0.052506 .
## GarageYrBlt       -30.2542      98.7897  -0.306         0.759481
## GarageCars      17995.5160    3728.8726   4.826 0.00000161685545 ***
## GarageArea          0.8644      12.8965   0.067         0.946577
## WoodDeckSF         23.5870      10.7496   2.194         0.028456 *
## OpenPorchSF       -12.7451      21.5523  -0.591         0.554419
## EnclosedPorch      10.7139      22.2777   0.481         0.630677
## X3SsnPorch          9.4727      40.2453   0.235         0.813969
## ScreenPorch        60.6462      21.6746   2.798         0.005243 **
## PoolArea          -60.2413      30.6315  -1.967         0.049508 *
## MiscVal            -3.1572       7.1943  -0.439         0.660867
## MoSold           -290.6116     453.4291  -0.641         0.521726
## YrSold           -106.8355     908.4801  -0.118         0.906410
## YearsOld                NA           NA      NA               NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37450 on 972 degrees of freedom
## Multiple R-squared:  0.8075, Adjusted R-squared:  0.8006
## F-statistic: 116.5 on 35 and 972 DF,  p-value: < 2.2e-16
```

```
# Pick up the most significant (***) parameters and rerun the model
model = lm(SalePrice~MSSubClass+LotArea+OverallQual+
             OverallCond+MasVnrArea+YearsOld,
           data=train)
summary(model)
```

```
##
## Call:
## lm(formula = SalePrice ~ MSSubClass + LotArea + OverallQual +
##     OverallCond + MasVnrArea + YearsOld, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -316503  -25139   -3237   20383  403781
```

```
##
## Coefficients:
##                Estimate  Std. Error t value         Pr(>|t|)
## (Intercept) -86271.1638  11884.3026  -7.259 0.00000000000078 ***
## MSSubClass    -156.3186     35.0377  -4.461 0.00000906357575 ***
## LotArea          1.5313      0.1764   8.680          < 2e-16 ***
## OverallQual  39892.8578   1388.7570  28.726          < 2e-16 ***
## OverallCond   3859.7187   1516.7666   2.545           0.0111 *
## MasVnrArea      73.1256      8.4394   8.665          < 2e-16 ***
## YearsOld      -264.0750     63.4673  -4.161 0.00003443677037 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45600 on 1001 degrees of freedom
## Multiple R-squared:  0.7062, Adjusted R-squared:  0.7044
## F-statistic:   401 on 6 and 1001 DF,  p-value: < 2.2e-16
```

```
preds = predict(model, newdata = test)



res = data.frame(test$SalePrice, round(preds,0))
names(res) = c("Actual", "Pred")
# Lets look at top 10 predictions and actual values
# just to visually compare
head(res,10)
```

```
##     Actual   Pred
## 7   307000 274918
## 20  139000 129485
## 47  239686 220663
## 48  249700 263395
## 50  127000 135736
## 61  158000 186107
## 63  202500 253652
## 75  107400  30944
## 88  164500 157133
## 138 171000 212885
```

```
# Lets calc Mean Squared Error
mse = sum((res$Actual-res$Pred)^2)/nrow(res)
mse
```

```
## [1] 1795965975
```

```
# Is the MSE any better than mean of train output (which would
# be our simplest/naiivest guess)
naiveMSE = sum((res$Actual-mean(train$SalePrice))^2)/nrow(res)
naiveMSE
```

```
## [1] 5580012526
```

```r
# Naive Mean Squared error is bigger than our MSE (not bad)
naiveMSE/mse
```

```
## [1] 3.10697
```