Kaggle Housing Price Project

Huidi Wang

March 13, 2017

Project description

The data used in this project are sourced from Kaggle Housing Project data with 1459 test and 1480 train observations. It shows the relationship between housing price and other thirteen related varibles, such as neighborhood, yearsold, building type, etc.. Our goal is to figure out how these measurements influence the reference viarible through linear regression, decision tree, random forest, and boosting methods.

1. Understand data and find out NAs

1) Understand all data

A summary of 81 variables

```
## 'data.frame':    1460 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : chr  "RL" "RL" "RL" "RL" ...
##  $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
##  $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
##  $ Street       : chr  "Pave" "Pave" "Pave" "Pave" ...
##  $ Alley        : chr  NA NA NA NA ...
##  $ LotShape     : chr  "Reg" "Reg" "IR1" "IR1" ...
##  $ LandContour  : chr  "Lvl" "Lvl" "Lvl" "Lvl" ...
##  $ Utilities    : chr  "AllPub" "AllPub" "AllPub" "AllPub" ...
##  $ LotConfig    : chr  "Inside" "FR2" "Inside" "Corner" ...
##  $ LandSlope    : chr  "Gtl" "Gtl" "Gtl" "Gtl" ...
##  $ Neighborhood : chr  "CollgCr" "Veenker" "CollgCr" "Crawfor" ...
##  $ Condition1   : chr  "Norm" "Feedr" "Norm" "Norm" ...
##  $ Condition2   : chr  "Norm" "Norm" "Norm" "Norm" ...
##  $ BldgType     : chr  "1Fam" "1Fam" "1Fam" "1Fam" ...
##  $ HouseStyle   : chr  "2Story" "1Story" "2Story" "2Story" ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
##  $ RoofStyle    : chr  "Gable" "Gable" "Gable" "Gable" ...
##  $ RoofMatl     : chr  "CompShg" "CompShg" "CompShg" "CompShg" ...
##  $ Exterior1st  : chr  "VinylSd" "MetalSd" "VinylSd" "Wd Sdng" ...
##  $ Exterior2nd  : chr  "VinylSd" "MetalSd" "VinylSd" "Wd Shng" ...
##  $ MasVnrType   : chr  "BrkFace" "None" "BrkFace" "None" ...
##  $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : chr  "Gd" "TA" "Gd" "TA" ...
##  $ ExterCond    : chr  "TA" "TA" "TA" "TA" ...
##  $ Foundation   : chr  "PConc" "CBlock" "PConc" "BrkTil" ...
##  $ BsmtQual     : chr  "Gd" "Gd" "Gd" "TA" ...
##  $ BsmtCond     : chr  "TA" "TA" "TA" "Gd" ...
##  $ BsmtExposure : chr  "No" "Gd" "Mn" "No" ...
##  $ BsmtFinType1 : chr  "GLQ" "ALQ" "GLQ" "ALQ" ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : chr  "Unf" "Unf" "Unf" "Unf" ...
##  $ BsmtFinSF2   : int  0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF    : int  150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
##  $ Heating      : chr  "GasA" "GasA" "GasA" "GasA" ...
##  $ HeatingQC    : chr  "Ex" "Ex" "Ex" "Gd" ...
##  $ CentralAir   : chr  "Y" "Y" "Y" "Y" ...
##  $ Electrical   : chr  "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...
##  $ 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  : chr  "Gd" "TA" "Gd" "Gd" ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : chr  "Typ" "Typ" "Typ" "Typ" ...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : chr  NA "TA" "TA" "Gd" ...
##  $ GarageType   : chr  "Attchd" "Attchd" "Attchd" "Detchd" ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
##  $ GarageFinish : chr  "RFn" "RFn" "RFn" "Unf" ...
##  $ GarageCars   : int  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : chr  "TA" "TA" "TA" "TA" ...
##  $ GarageCond   : chr  "TA" "TA" "TA" "TA" ...
##  $ PavedDrive   : chr  "Y" "Y" "Y" "Y" ...
##  $ WoodDeckSF   : int  0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF  : int  61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch: int  0 0 0 272 0 0 0 228 205 0 ...
##  $ 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       : chr  NA NA NA NA ...
##  $ Fence        : chr  NA NA NA NA ...
##  $ MiscFeature  : chr  NA NA NA NA ...
##  $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
##  $ SaleType     : chr  "WD" "WD" "WD" "WD" ...
##  $ SaleCondition: chr  "Normal" "Normal" "Normal" "Abnorml" ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

Sample of head of train data

```
##   Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape
## 1  1         60       RL          65    8450   Pave  <NA>      Reg
## 2  2         20       RL          80    9600   Pave  <NA>      Reg
## 3  3         60       RL          68   11250   Pave  <NA>      IR1
## 4  4         70       RL          60    9550   Pave  <NA>      IR1
## 5  5         60       RL          84   14260   Pave  <NA>      IR1
## 6  6         50       RL          85   14115   Pave  <NA>      IR1
```

2)Check how many NAs in each feature in order

```
##        PoolQC   MiscFeature         Alley         Fence   FireplaceQu 
##          1453          1406          1369          1179           690 
##   LotFrontage    GarageType   GarageYrBlt  GarageFinish    GarageQual 
##           259            81            81            81            81 
##    GarageCond  BsmtExposure  BsmtFinType2      BsmtQual      BsmtCond 
##            81            38            38            37            37 
##  BsmtFinType1    MasVnrType    MasVnrArea    Electrical            Id 
##            37             8             8             1             0 
##    MSSubClass      MSZoning       LotArea        Street      LotShape 
##             0             0             0             0             0 
##   LandContour     Utilities     LotConfig     LandSlope  Neighborhood 
##             0             0             0             0             0 
##    Condition1    Condition2      BldgType    HouseStyle   OverallQual 
##             0             0             0             0             0 
##   OverallCond     YearBuilt  YearRemodAdd     RoofStyle      RoofMatl 
##             0             0             0             0             0 
##   Exterior1st   Exterior2nd     ExterQual     ExterCond    Foundation 
##             0             0             0             0             0 
##    BsmtFinSF1    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating 
##             0             0             0             0             0 
##     HeatingQC    CentralAir      X1stFlrSF     X2ndFlrSF  LowQualFinSF 
##             0             0             0             0             0 
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath 
##             0             0             0             0             0 
##  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional 
##             0             0             0             0             0 
##    Fireplaces    GarageCars    GarageArea    PavedDrive    WoodDeckSF 
##             0             0             0             0             0 
##   OpenPorchSF EnclosedPorch    X3SsnPorch   ScreenPorch      PoolArea 
##             0             0             0             0             0 
##       MiscVal        MoSold        YrSold      SaleType SaleCondition 
##             0             0             0             0             0 
##     SalePrice 
##             0
```

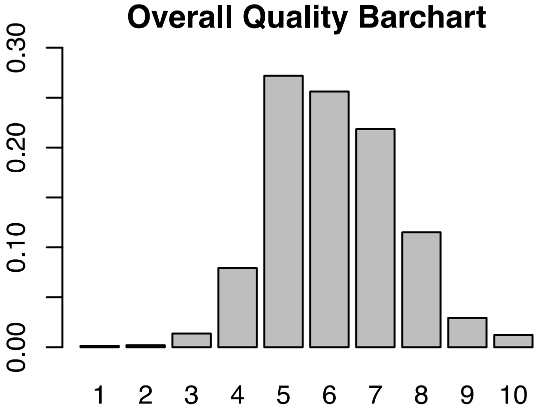
Percentage of NAs in all train dataset

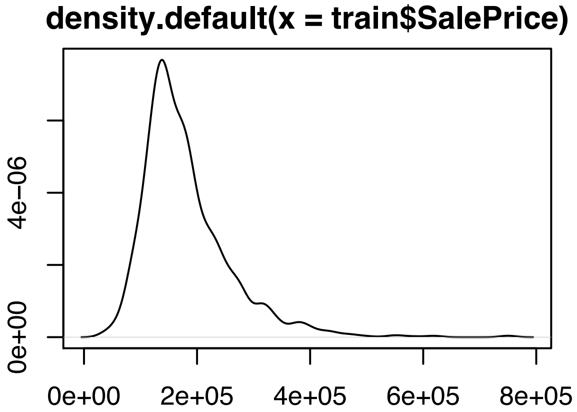
```
## [1] 0.05889565
```

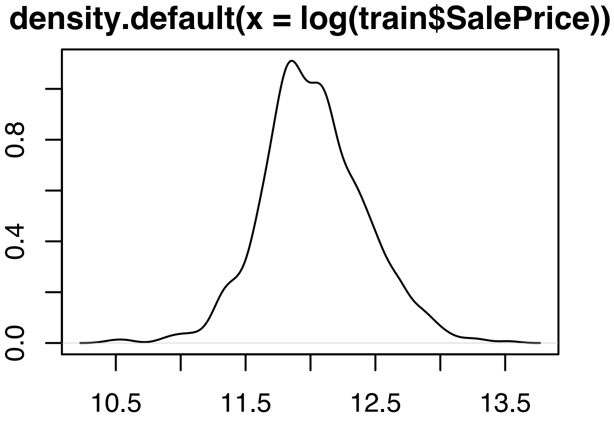
3) To deal with missing values: (1) Actually missing too many values: delete them (2) Some missing values are meaningful (like BsmtUnfSF~BsmtExposure, BsmtFinType2, BsmtQual, BsmtCond, BsmtFinType1) (3) Simple missing might due to operation or data transfer –> we could use different ways to impute, such as mean, or median (4) More advanced way is to use model based: using other features to predict the missing value (MICE)

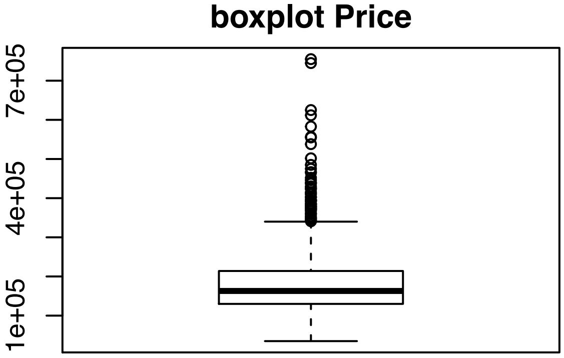
2. Analyze each variable both categorical (barchart) and continuous (plot or frequency)

1) Show relationship using barplot and boxplot

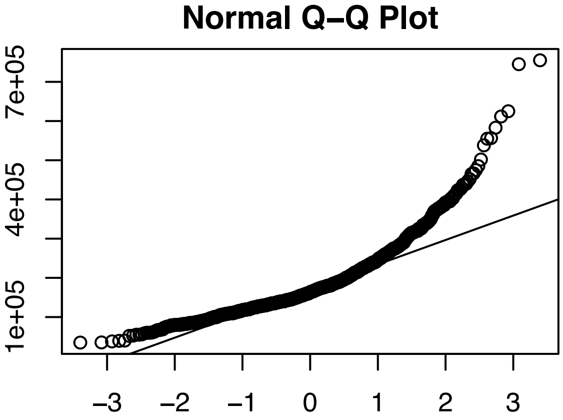


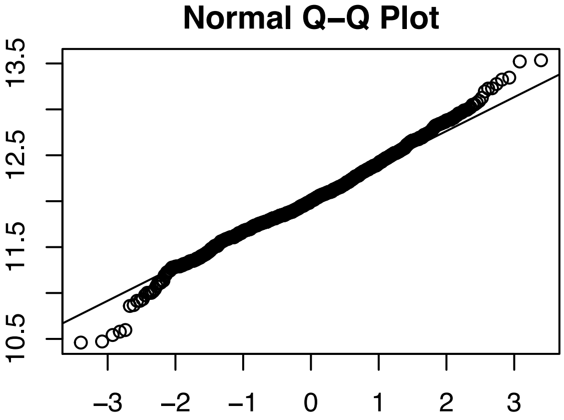






2) Check Normality and use log(SalePrice) for following all analysis





3. Find out Important Features

1) Combine some correlated features.

```r
# 1stflr+2ndflr+lowqualsf+GrLivArea = All_Liv_Area
train_noNA$AllSF <- with(train_noNA, X1stFlrSF+X2ndFlrSF+GrLivArea + TotalBsmtSF)
# Total number of bathrooms
train_noNA$TotalBath <- with(train_noNA, BsmtFullBath + 0.5 * BsmtHalfBath + FullBath + 0.5 * HalfBath)
#remove unnesessary features
drops=c("Id", "BsmtFullBath" , "BsmtHalfBath", "FullBath", "HalfBath", "X1stFlrSF","X2ndFlrSF","GrLivAre
train_noNA=train_noNA[,!names(train_noNA)%in%drops]
raw_0=train_noNA
```

2) Backward selection: ignore the biggest p-value after doing linear regression

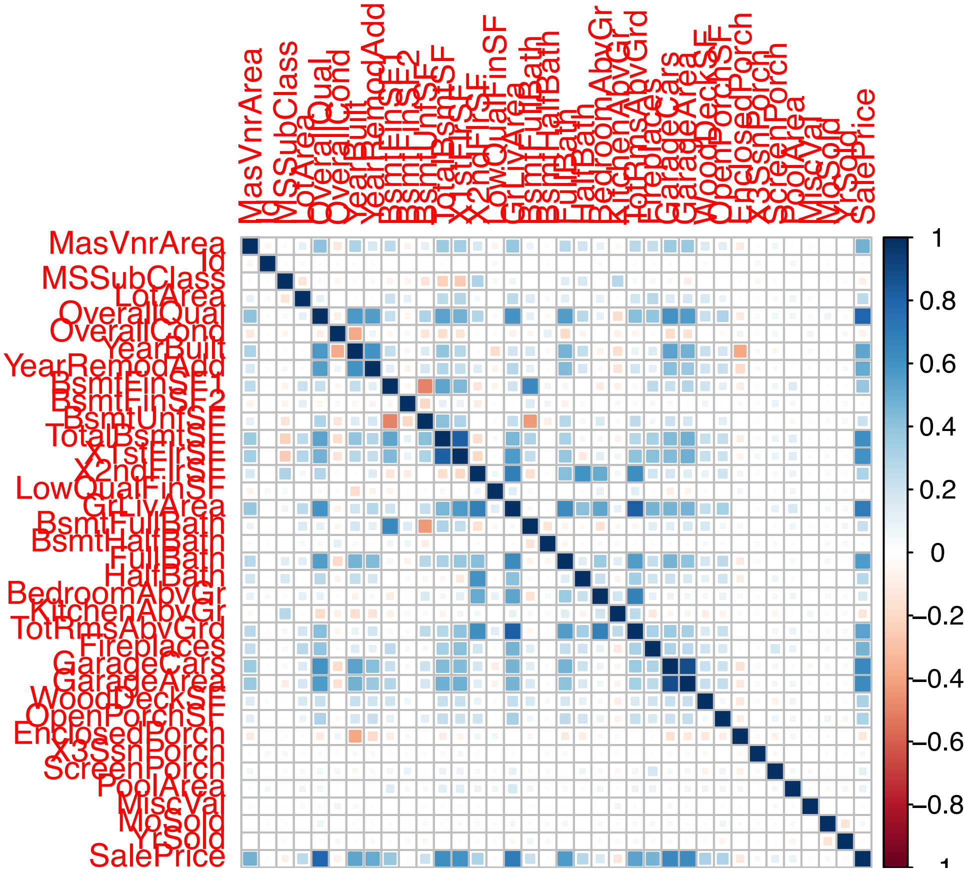
```
##
## Call:
## lm(formula = SalePrice ~ ., data = raw_0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69841 -0.04675  0.00171  0.05303  0.69841
##
## Coefficients: (2 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.949e+00  4.772e+00   1.456 0.145595
## BsmtExposureGd         2.733e-02  1.374e-02   1.989 0.046963 *
## BsmtExposureMn        -8.326e-03  1.377e-02  -0.605 0.545598
## BsmtExposureNo        -1.233e-02  9.969e-03  -1.236 0.216514
## BsmtExposureUnf       -5.101e-02  1.064e-01  -0.479 0.631779
## BsmtFinType2BLQ       -6.916e-02  3.445e-02  -2.007 0.044930 *
## BsmtFinType2GLQ       -2.729e-03  4.264e-02  -0.064 0.948978
## BsmtFinType2LwQ       -3.651e-02  3.353e-02  -1.089 0.276483
## BsmtFinType2Rec       -2.547e-02  3.231e-02  -0.788 0.430690
## BsmtFinType2Unf       -1.331e-02  3.424e-02  -0.389 0.697577
## BsmtQualFa            -3.506e-02  2.874e-02  -1.220 0.222727
## BsmtQualGd            -2.782e-02  1.503e-02  -1.850 0.064519 .
## BsmtQualTA            -3.479e-02  1.862e-02  -1.868 0.061928 .
## BsmtQualUnf            5.286e-02  1.164e-01   0.454 0.649950
## BsmtCondGd             1.848e-02  2.391e-02   0.773 0.439871
## BsmtCondPo             3.349e-01  1.362e-01   2.459 0.014081 *
## BsmtCondTA             2.088e-02  1.906e-02   1.095 0.273677
## BsmtCondUnf                   NA         NA      NA       NA
## BsmtFinType1BLQ       -5.175e-03  1.258e-02  -0.411 0.680940
## BsmtFinType1GLQ        9.384e-03  1.145e-02   0.819 0.412699
## BsmtFinType1LwQ       -2.392e-02  1.692e-02  -1.413 0.157826
## BsmtFinType1Rec       -6.425e-03  1.353e-02  -0.475 0.635068
## BsmtFinType1Unf       -1.431e-02  1.330e-02  -1.075 0.282421
## MasVnrTypeBrkFace      3.456e-02  3.094e-02   1.117 0.264132
## MasVnrTypeNone         2.791e-02  3.117e-02   0.895 0.370735
## MasVnrTypeStone        4.121e-02  3.271e-02   1.260 0.207969
## MasVnrArea             8.724e-06  2.619e-05   0.333 0.739122
## ElectricalFuseF        7.789e-04  2.642e-02   0.029 0.976488
## ElectricalFuseP       -4.084e-02  7.804e-02  -0.523 0.600881
## ElectricalMix         -2.038e-01  1.826e-01  -1.116 0.264575
## ElectricalSBrkr       -1.426e-02  1.346e-02  -1.060 0.289444
## MSSubClass            -3.684e-04  3.766e-04  -0.978 0.328188
## MSZoningFV            4.438e-01  5.426e-02   8.180 6.93e-16 ***
## MSZoningRH            4.153e-01  5.402e-02   7.687 3.04e-14 ***
## MSZoningRL            4.179e-01  4.628e-02   9.030  < 2e-16 ***
## MSZoningRM            3.750e-01  4.336e-02   8.647  < 2e-16 ***
## LotArea               2.981e-06  4.909e-07   6.071 1.68e-09 ***
## StreetPave            1.154e-01  5.443e-02   2.120 0.034184 *
## LotShapeIR2           2.791e-02  1.905e-02   1.465 0.143277
## LotShapeIR3           1.605e-02  3.995e-02   0.402 0.687960
## LotShapeReg           5.184e-03  7.340e-03   0.706 0.480136
## LandContourHLS        2.741e-02  2.341e-02   1.171 0.241870
## LandContourLow       -2.430e-02  2.912e-02  -0.834 0.404292
## LandContourLvl        2.596e-02  1.678e-02   1.547 0.122181
## UtilitiesNoSeWa      -2.243e-01  1.175e-01  -1.909 0.056486 .
## LotConfigCulDSac      2.670e-02  1.461e-02   1.827 0.067946 .
## LotConfigFR2         -4.051e-02  1.818e-02  -2.228 0.026041 *
## LotConfigFR3         -9.698e-02  5.761e-02  -1.683 0.092555 .
## LotConfigInside      -1.392e-02  7.973e-03  -1.746 0.081027 .
## LandSlopeMod          3.102e-02  1.808e-02   1.716 0.086372 .
## LandSlopeSev         -1.954e-01  5.175e-02  -3.775 0.000167 ***
## NeighborhoodBlueste  -4.363e-02  8.701e-02  -0.501 0.616147
## NeighborhoodBrDale   -5.639e-02  4.980e-02  -1.132 0.257737
## NeighborhoodBrkSide   1.542e-02  4.269e-02   0.361 0.718033
## NeighborhoodClearCr   1.799e-02  4.192e-02   0.429 0.667808
## NeighborhoodCollgCr  -1.648e-02  3.263e-02  -0.505 0.613729
## NeighborhoodCrawfor   1.121e-01  3.870e-02   2.897 0.003838 **
## NeighborhoodEdwards  -8.154e-02  3.607e-02  -2.261 0.023960 *
## NeighborhoodGilbert  -8.024e-03  3.502e-02  -0.229 0.818809
## NeighborhoodIDOTRR   -2.400e-02  4.850e-02  -0.495 0.620719
## NeighborhoodMeadowV  -1.569e-01  5.055e-02  -3.104 0.001952 **
## NeighborhoodMitchel  -5.812e-02  3.692e-02  -1.574 0.115675
## NeighborhoodNAmes    -3.237e-02  3.529e-02  -0.917 0.359190
## NeighborhoodNoRidge   3.853e-02  3.807e-02   1.012 0.311757
## NeighborhoodNPkVill  -6.299e-03  6.346e-02  -0.099 0.920951
## NeighborhoodNridgHt   7.765e-02  3.351e-02   2.317 0.020653 *
## NeighborhoodNWAmes   -3.885e-02  3.639e-02  -1.067 0.285963
## NeighborhoodOldTown  -5.174e-02  4.350e-02  -1.189 0.234520
## NeighborhoodSawyer   -2.164e-02  3.687e-02  -0.587 0.557263
## NeighborhoodSawyerW  -5.115e-03  3.527e-02  -0.145 0.884699
## NeighborhoodSomerst   2.298e-02  4.083e-02   0.563 0.573706
## NeighborhoodStoneBr   1.337e-01  3.764e-02   3.553 0.000395 ***
## NeighborhoodSWISU     1.581e-03  4.373e-02   0.036 0.971170
## NeighborhoodTimber    3.832e-03  3.694e-02   0.104 0.917396
## NeighborhoodVeenker   4.928e-02  4.764e-02   1.034 0.301136
## Condition1Feedr       2.642e-02  2.228e-02   1.186 0.235951
## Condition1Norm        7.559e-02  1.845e-02   4.096 4.47e-05 ***
## Condition1PosA        4.275e-02  4.507e-02   0.948 0.343087
## Condition1PosN        7.951e-02  3.347e-02   2.376 0.017673 *
## Condition1RRAe       -4.346e-02  4.090e-02  -1.063 0.288174
## Condition1RRAn        3.223e-02  3.086e-02   1.044 0.296518
## Condition1RRNe        8.447e-03  8.039e-02   0.105 0.916334
## Condition1RRNn        8.473e-02  5.802e-02   1.460 0.144455
## Condition2Feedr       1.173e-01  1.013e-01   1.158 0.247198
## Condition2Norm        5.494e-02  8.652e-02   0.635 0.525538
## Condition2PosA        2.427e-01  1.667e-01   1.456 0.145702
## Condition2PosN       -8.144e-01  1.218e-01  -6.685 3.48e-11 ***
## Condition2RRAe       -5.405e-01  2.067e-01  -2.615 0.009039 ** 
## Condition2RRAn       -2.697e-02  1.401e-01  -0.192 0.847422    
## Condition2RRNn        2.411e-02  1.197e-01   0.201 0.840451    
## BldgType2fmCon        4.427e-02  5.677e-02   0.780 0.435605    
## BldgTypeDuplex       -9.866e-03  3.305e-02  -0.299 0.765344    
## BldgTypeTwnhs        -5.644e-02  4.510e-02  -1.251 0.211050    
## BldgTypeTwnhsE       -1.119e-02  4.063e-02  -0.276 0.782936    
## HouseStyle1.5Unf      7.380e-03  3.411e-02   0.216 0.828773    
## HouseStyle1Story     -2.687e-02  1.679e-02  -1.600 0.109803    
## HouseStyle2.5Fin     -5.839e-02  5.324e-02  -1.097 0.272950    
## HouseStyle2.5Unf      5.299e-02  4.089e-02   1.296 0.195157    
## HouseStyle2Story     -1.949e-02  1.466e-02  -1.329 0.184090    
## HouseStyleSFoyer     -1.745e-02  2.774e-02  -0.629 0.529407    
## HouseStyleSLvl       -1.350e-03  2.376e-02  -0.057 0.954720    
## OverallQual           4.443e-02  4.575e-03   9.712  < 2e-16 ***
## OverallCond           3.722e-02  3.948e-03   9.426  < 2e-16 ***
## YearBuilt             1.730e-03  3.310e-04   5.226 2.03e-07 ***
## YearRemodAdd          7.671e-04  2.478e-04   3.095 0.002010 ** 
## RoofStyleGable       -1.845e-02  8.323e-02  -0.222 0.824651    
## RoofStyleGambrel     -2.959e-03  9.105e-02  -0.032 0.974080    
## RoofStyleHip         -1.384e-02  8.348e-02  -0.166 0.868396    
## RoofStyleMansard      5.837e-02  9.690e-02   0.602 0.547079    
## RoofStyleShed         4.781e-01  1.577e-01   3.032 0.002481 ** 
## RoofMatlCompShg       2.586e+00  1.492e-01  17.334  < 2e-16 ***
## RoofMatlMembran       2.980e+00  2.172e-01  13.723  < 2e-16 ***
## RoofMatlMetal         2.836e+00  2.127e-01  13.331  < 2e-16 ***
## RoofMatlRoll          2.593e+00  1.874e-01  13.833  < 2e-16 ***
## RoofMatlTar&Grv       2.611e+00  1.714e-01  15.240  < 2e-16 ***
## RoofMatlWdShake       2.512e+00  1.651e-01  15.221  < 2e-16 ***
## RoofMatlWdShngl       2.684e+00  1.544e-01  17.388  < 2e-16 ***
## Exterior1stAsphShn    3.398e-02  1.508e-01   0.225 0.821729    
## Exterior1stBrkComm   -1.866e-01  1.253e-01  -1.490 0.136511    
## Exterior1stBrkFace    1.076e-01  5.649e-02   1.906 0.056938 .  
## Exterior1stCBlock    -5.132e-02  1.235e-01  -0.416 0.677708    
## Exterior1stCemntBd   -6.784e-02  8.535e-02  -0.795 0.426878    
## Exterior1stHdBoard    1.685e-02  5.714e-02   0.295 0.768175    
## Exterior1stImStucc    1.676e-03  1.255e-01   0.013 0.989349    
## Exterior1stMetalSd    5.985e-02  6.500e-02   0.921 0.357360    
## Exterior1stPlywood    1.697e-02  5.649e-02   0.300 0.763932    
## Exterior1stStone      9.087e-02  1.091e-01   0.833 0.405136    
## Exterior1stStucco     4.925e-02  6.206e-02   0.794 0.427548    
## Exterior1stVinylSd    1.857e-02  5.892e-02   0.315 0.752642    
## Exterior1stWd Sdng   -1.480e-02  5.451e-02  -0.272 0.786010    
## Exterior1stWdShing    1.896e-02  5.892e-02   0.322 0.747647    
## Exterior2ndAsphShn   -6.995e-03  9.970e-02  -0.070 0.944076    
## Exterior2ndBrk Cmn    1.891e-02  9.087e-02   0.208 0.835151    
## Exterior2ndBrkFace   -5.069e-02  5.888e-02  -0.861 0.389460    
## Exterior2ndCBlock            NA         NA      NA       NA    
## Exterior2ndCmentBd    1.115e-01  8.416e-02   1.325 0.185440    
## Exterior2ndHdBoard   -5.151e-03  5.531e-02  -0.093 0.925823    
## Exterior2ndImStucc    1.128e-02  6.376e-02   0.177 0.859534    
## Exterior2ndMetalSd   -1.923e-02  6.365e-02  -0.302 0.762658    
## Exterior2ndOther     -1.015e-01  1.244e-01  -0.816 0.414761    
## Exterior2ndPlywood   -1.065e-03  5.360e-02  -0.020 0.984146
## Exterior2ndStone     -8.227e-02  7.702e-02  -1.068 0.285598
## Exterior2ndStucco    -1.461e-02  5.997e-02  -0.244 0.807547
## Exterior2ndVinylSd    1.506e-02  5.715e-02   0.263 0.792222
## Exterior2ndWd Sdng    3.450e-02  5.292e-02   0.652 0.514572
## Exterior2ndWd Shng   -1.214e-02  5.483e-02  -0.221 0.824852
## ExterQualFa           2.658e-02  4.945e-02   0.538 0.591001
## ExterQualGd           9.121e-03  2.192e-02   0.416 0.677355
## ExterQualTA           1.440e-02  2.423e-02   0.594 0.552454
## ExterCondFa          -8.531e-02  8.227e-02  -1.037 0.300008
## ExterCondGd          -5.870e-02  7.874e-02  -0.746 0.456087
## ExterCondPo          -8.880e-02  1.435e-01  -0.619 0.536204
## ExterCondTA          -3.971e-02  7.850e-02  -0.506 0.613090
## FoundationCBlock      2.367e-02  1.430e-02   1.655 0.098195 .
## FoundationPConc       3.973e-02  1.546e-02   2.570 0.010293 *
## FoundationSlab       -3.530e-02  4.555e-02  -0.775 0.438514
## FoundationStone       1.270e-01  4.912e-02   2.586 0.009831 **
## FoundationWood       -1.192e-01  6.654e-02  -1.792 0.073363 .
## BsmtFinSF1            3.005e-05  2.375e-05   1.265 0.206101
## BsmtFinSF2            2.720e-05  4.020e-05   0.677 0.498792
## BsmtUnfSF            -2.890e-05  2.247e-05  -1.286 0.198751
## HeatingGasA           1.595e-01  1.158e-01   1.378 0.168512
## HeatingGasW           2.224e-01  1.190e-01   1.868 0.062021 .
## HeatingGrav           8.587e-03  1.254e-01   0.068 0.945428
## HeatingOthW           1.407e-01  1.429e-01   0.984 0.325133
## HeatingWall           2.606e-01  1.341e-01   1.943 0.052197 .
## HeatingQCFa          -2.304e-02  2.131e-02  -1.081 0.279884
## HeatingQCGd          -2.143e-02  9.418e-03  -2.275 0.023065 *
## HeatingQCPo          -1.086e-01  1.224e-01  -0.887 0.375247
## HeatingQCTA          -3.316e-02  9.386e-03  -3.533 0.000426 ***
## CentralAirY           6.799e-02  1.764e-02   3.855 0.000121 ***
## LowQualFinSF          8.422e-05  8.038e-05   1.048 0.294918
## BedroomAbvGr          3.620e-03  6.079e-03   0.596 0.551575
## KitchenAbvGr         -4.085e-02  2.542e-02  -1.607 0.108334
## KitchenQualFa        -5.863e-02  2.821e-02  -2.079 0.037858 *
## KitchenQualGd        -6.512e-02  1.552e-02  -4.195 2.92e-05 ***
## KitchenQualTA        -6.585e-02  1.763e-02  -3.735 0.000197 ***
## TotRmsAbvGrd          5.498e-03  4.292e-03   1.281 0.200453
## FunctionalMaj2       -2.284e-01  6.574e-02  -3.474 0.000530 ***
## FunctionalMin1        3.774e-02  3.912e-02   0.965 0.334935
## FunctionalMin2        3.566e-02  3.884e-02   0.918 0.358666
## FunctionalMod        -5.821e-02  4.757e-02  -1.224 0.221298
## FunctionalSev        -2.784e-01  1.264e-01  -2.203 0.027780 *
## FunctionalTyp         6.962e-02  3.369e-02   2.067 0.038986 *
## Fireplaces            2.517e-02  6.033e-03   4.172 3.23e-05 ***
## GarageCars            2.466e-02  9.773e-03   2.524 0.011739 *
## GarageArea            1.225e-04  3.368e-05   3.637 0.000287 ***
## PavedDriveP           1.620e-02  2.456e-02   0.660 0.509514
## PavedDriveY           2.296e-02  1.545e-02   1.486 0.137459
## WoodDeckSF            9.431e-05  2.642e-05   3.569 0.000372 ***
## OpenPorchSF           6.601e-05  5.211e-05   1.267 0.205481
## EnclosedPorch         1.238e-04  5.649e-05   2.191 0.028649 *
## X3SsnPorch            1.649e-04  1.017e-04   1.621 0.105376
## ScreenPorch           2.726e-04  5.528e-05   4.932 9.23e-07 ***
## PoolArea               1.525e-04  8.192e-05   1.862 0.062846 .  
## MiscVal               -2.967e-07  6.470e-06  -0.046 0.963429    
## MoSold                -8.104e-04  1.117e-03  -0.725 0.468427    
## YrSold                -2.231e-03  2.347e-03  -0.951 0.341977    
## SaleTypeCon            8.586e-02  8.080e-02   1.063 0.288152    
## SaleTypeConLD          1.325e-01  4.406e-02   3.007 0.002688 ** 
## SaleTypeConLI         -4.138e-02  5.232e-02  -0.791 0.429231    
## SaleTypeConLw          5.261e-03  5.530e-02   0.095 0.924226    
## SaleTypeCWD            6.346e-02  5.915e-02   1.073 0.283558    
## SaleTypeNew            7.262e-02  7.098e-02   1.023 0.306508    
## SaleTypeOth            6.095e-02  6.618e-02   0.921 0.357297    
## SaleTypeWD            -2.292e-02  1.911e-02  -1.199 0.230618    
## SaleConditionAdjLand  1.023e-01  6.586e-02   1.553 0.120562    
## SaleConditionAlloca   7.890e-02  3.874e-02   2.037 0.041884 *  
## SaleConditionFamily   1.588e-02  2.784e-02   0.570 0.568498    
## SaleConditionNormal   6.851e-02  1.316e-02   5.205 2.26e-07 ***
## SaleConditionPartial  2.060e-02  6.838e-02   0.301 0.763233    
## AllSF                  1.125e-04  9.026e-06  12.464  < 2e-16 ***
## TotalBath              2.241e-02  7.057e-03   3.175 0.001534 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1049 on 1249 degrees of freedom
## Multiple R-squared:  0.941,  Adjusted R-squared:  0.9311 
## F-statistic: 94.84 on 210 and 1249 DF,  p-value: < 2.2e-16
```

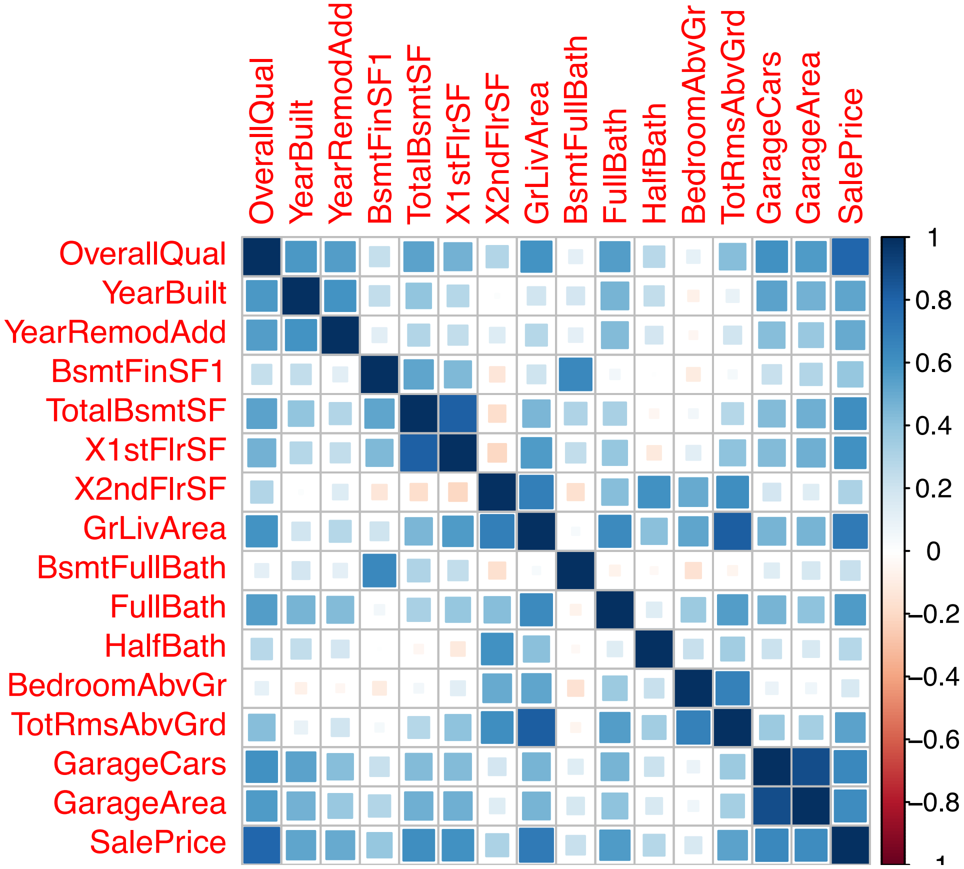
So relative weak features would be c(“BsmtCond”,“MasVnrType”,“Electrical”,“Id”,“MSSubClass”,“LotShape”, “HouseStyle”,“Exterior1st”,“Exterior2nd”,“Foundation”,“CentralAirY”,“PavedDrive”,“PavedDrive”,“EnclosedPorch”, “MiscVal”)

3) Using Lasso to find most predictable features, which will be described in next Prediction part.

4) Seperate categorical data and continuous data, get to explore relationship between each feature and outcome

–Numerical Data: Use corrplot to pick stronger continuous feature (Correlation)


MasVnrArea
Id
MSSubClass
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF
X1stFlrSF
X2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorch
X3SsnPorch
ScreenPorch
PoolArea
MiscVal
MoSold
YrSold
SalePrice
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1



```
#Basically, the darker blue square shows stronger relationship, so the stronger features might be:
numer_var
```

```
##  [1] "OverallQual"  "YearBuilt"    "YearRemodAdd" "BsmtFinSF1"
##  [5] "TotalBsmtSF"  "X1stFlrSF"    "X2ndFlrSF"    "GrLivArea"
##  [9] "BsmtFullBath" "FullBath"     "HalfBath"     "BedroomAbvGr"
## [13] "TotRmsAbvGrd" "GarageCars"   "GarageArea"   "SalePrice"
```

—Categorical Data: Using Tabplot

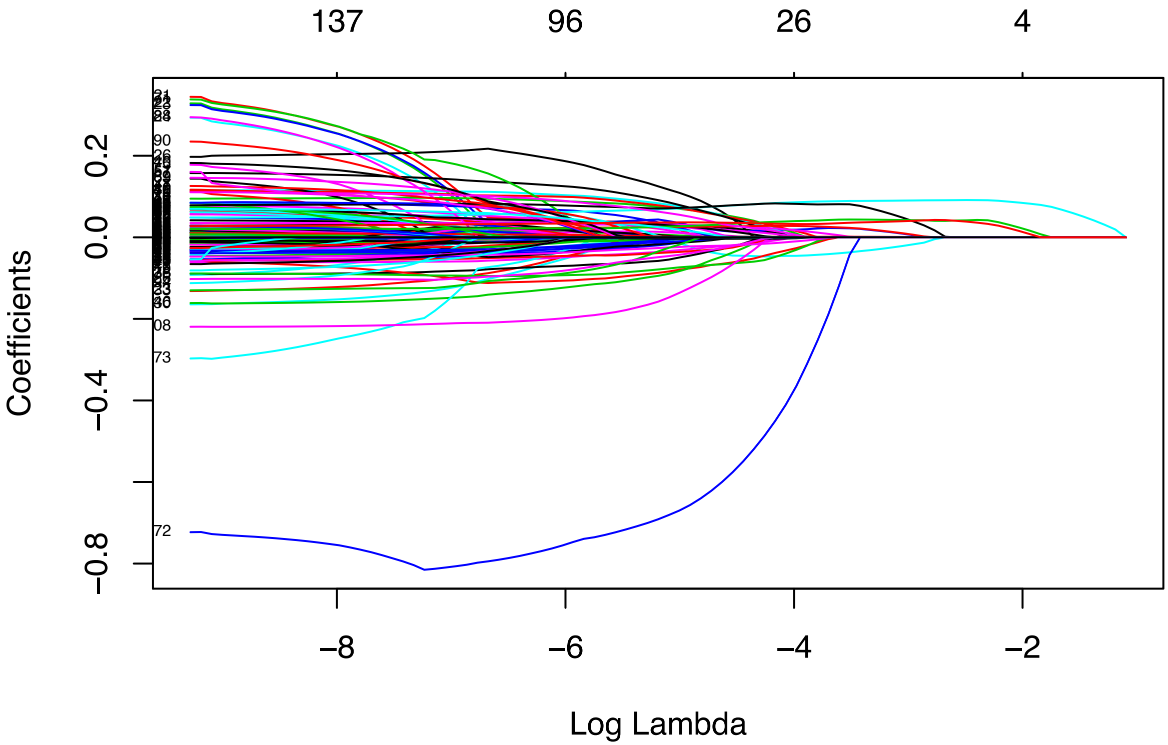
```
#for (i in 1:5) {
#  plot(tableplot(train[,categ_var], select = c(1, ((i - 1) * 5 + 1):(i * 5)),
#                  nBins = 100, plot = FALSE), fontsize = 12)
#}
#train$SalePrice=log(train$SalePrice)
```

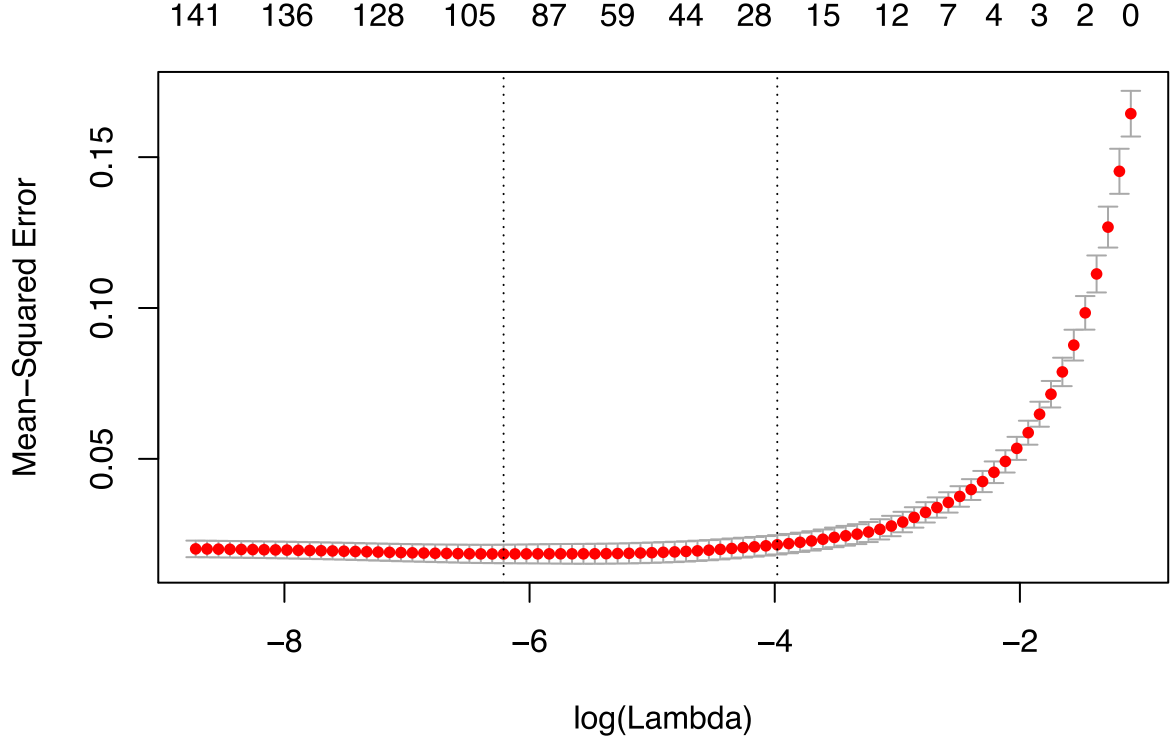
3. Prediction

1) Ridge and Lasso
   VIF is calculated for each feature and select all vif(feature)>10. Showing that lots of features exist correlation with other features. I'll use Lasso and Ridge to eliminate it and predict results.

```
##      BsmtExposureUnf      BsmtFinType2Unf           BsmtQualTA
##             37.85267             14.99301             10.57571
##          BsmtQualUnf          MSZoningFV           MSZoningRL
##             39.39303            15.59102            44.74239 
##           MSZoningRM  NeighborhoodCollgCr  NeighborhoodEdwards 
##             29.88276            12.26569            10.14276 
##    NeighborhoodNAmes  NeighborhoodOldTown  NeighborhoodSomerst 
##             19.62115            16.10986            11.12724 
##       RoofStyleGable         RoofStyleHip      RoofMatlCompShg 
##            149.74416           139.21273            43.46842 
##      RoofMatlTar&Grv      RoofMatlWdShake      RoofMatlWdShngl 
##             24.96338            10.01503            10.93675 
##          ExterQualGd          ExterQualTA          ExterCondFa 
##             13.30469            17.29148            16.48221 
##          ExterCondGd          ExterCondTA          HeatingGasA 
##             72.91008            86.34043            34.93064 
##          HeatingGasW          SaleTypeNew SaleConditionPartial 
##             20.80051            49.75440            47.28757 
##                AllSF 
##             12.47341
```

Using lambda.1se to predict





Calculate sum of square error and RMSE

```
## [1] 14.20269
```

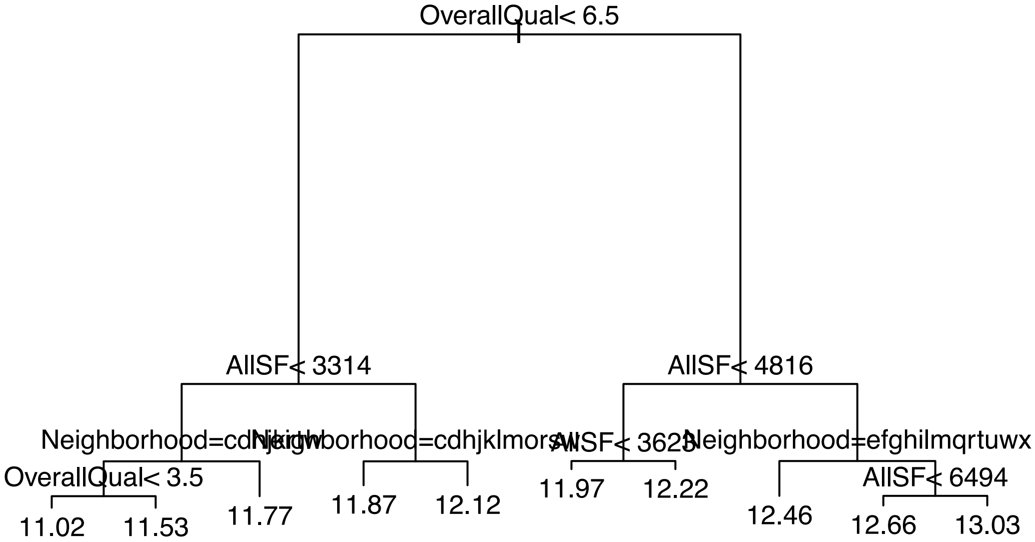
```
## [1] 0.1798677
```

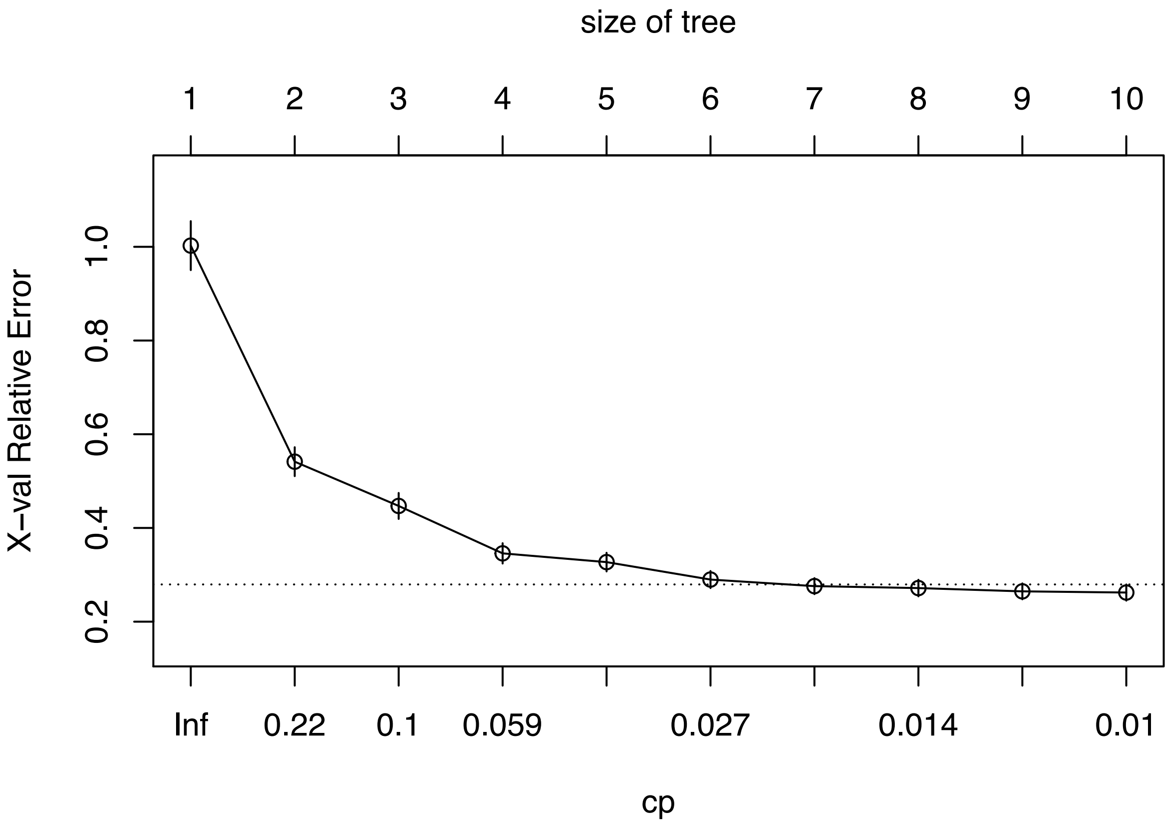
Find out strong features from Lasso method:

```
## 154 x 1 sparse Matrix of class "dgCMatrix"
##                                 1
## (Intercept)           6.259161e+00
## (Intercept)           .
## BsmtExposureGd        4.811460e-03
## BsmtExposureMn        .
## BsmtExposureNo        .
## BsmtExposureUnf       .
## BsmtFinType2BLQ       .
## BsmtFinType2GLQ       .
## BsmtFinType2LwQ       .
## BsmtFinType2Rec       .
## BsmtFinType2Unf       .
## BsmtQualFa            .
## BsmtQualGd            .
## BsmtQualTA            .
## BsmtQualUnf           .
## BsmtFinType1BLQ       .
## BsmtFinType1GLQ       .
## BsmtFinType1LwQ       .
## BsmtFinType1Rec       .
## BsmtFinType1Unf       .
## MasVnrArea            .
## MSZoningFV            .
## MSZoningRH            .
## MSZoningRL            1.876535e-02
## MSZoningRM           -4.541981e-02
## LotArea               7.332439e-07
## StreetPave            .
## LandContourHLS        .
## LandContourLow        .
## LandContourLvl        .
## UtilitiesNoSeWa       .
## LotConfigCulDSac      .
## LotConfigFR2          .
## LotConfigFR3          .
## LotConfigInside       .
## LandSlopeMod          .
## LandSlopeSev          .
## NeighborhoodBlueste   .
## NeighborhoodBrDale    .
## NeighborhoodBrkSide   .
## NeighborhoodClearCr   9.129175e-03
## NeighborhoodCollgCr   .
## NeighborhoodCrawfor   2.773373e-03
## NeighborhoodEdwards   .
## NeighborhoodGilbert   .
## NeighborhoodIDOTRR   -2.641041e-02
## NeighborhoodMeadowV   .
## NeighborhoodMitchel   .
## NeighborhoodNAmes     .
## NeighborhoodNoRidge   .
## NeighborhoodNPkVill   .
## NeighborhoodNridgHt   1.520541e-02
## NeighborhoodNWAmes    .
## NeighborhoodOldTown   .
## NeighborhoodSawyer    .
## NeighborhoodSawyerW   .
## NeighborhoodSomerst   .
## NeighborhoodStoneBr   .
## NeighborhoodSWISU     .
## NeighborhoodTimber    .
## NeighborhoodVeenker   .
## Condition1Feedr       .
## Condition1Norm        1.811148e-03
## Condition1PosA        .
## Condition1PosN        .
## Condition1RRAe        .
## Condition1RRAn        .
## Condition1RRNe        .
## Condition1RRNn        .
## Condition2Feedr       .
## Condition2Norm        .
## Condition2PosA        .
## Condition2PosN       -3.646675e-01
## Condition2RRAe        .
## Condition2RRAn        .
## Condition2RRNn        .
## BldgType2fmCon        .
## BldgTypeDuplex        .
## BldgTypeTwnhs        -3.061966e-02
## BldgTypeTwnhsE        .
## OverallQual           8.599476e-02
## OverallCond           1.722321e-02
## YearBuilt             9.696644e-04
## YearRemodAdd          1.247632e-03
## RoofStyleGable        .
## RoofStyleGambrel      .
## RoofStyleHip          .
## RoofStyleMansard      .
## RoofStyleShed         .
## RoofMatlCompShg       .
## RoofMatlMembran       .
## RoofMatlMetal         .
## RoofMatlRoll          .
## RoofMatlTar&Grv       .
## RoofMatlWdShake       .
## RoofMatlWdShngl       .
## ExterQualFa           .
## ExterQualGd           .
## ExterQualTA           .
## ExterCondFa           .
## ExterCondGd           .
## ExterCondPo           .
## ExterCondTA           .
## BsmtFinSF1            5.696794e-05
## BsmtFinSF2            .
## BsmtUnfSF             .
## HeatingGasA           .
## HeatingGasW           .
## HeatingGrav           .
## HeatingOthW           .
## HeatingWall           .
## HeatingQCFa           .
## HeatingQCGd           .
## HeatingQCPo           .
## HeatingQCTA          -9.795375e-03
## CentralAirY           8.197634e-02
## LowQualFinSF          .
## BedroomAbvGr          .
## KitchenAbvGr          .
## KitchenQualFa         .
## KitchenQualGd         .
## KitchenQualTA        -2.265578e-03
## TotRmsAbvGrd          .
## FunctionalMaj2        .
## FunctionalMin1        .
## FunctionalMin2        .
## FunctionalMod         .
## FunctionalSev         .
## FunctionalTyp         .
## Fireplaces            2.786946e-02
## GarageCars            3.985903e-02
## GarageArea            1.067005e-04
## WoodDeckSF            .
## OpenPorchSF           .
## X3SsnPorch            .
## ScreenPorch           .
## PoolArea              .
## MoSold                .
## YrSold                .
## SaleTypeCon           .
## SaleTypeConLD         .
## SaleTypeConLI         .
## SaleTypeConLw         .
## SaleTypeCWD           .
## SaleTypeNew           .
## SaleTypeOth           .
## SaleTypeWD            .
## SaleConditionAdjLand  .
## SaleConditionAlloca   .
## SaleConditionFamily   .
## SaleConditionNormal   .
## SaleConditionPartial  5.934660e-03
## AllSF                 1.061840e-04
## TotalBath             3.252878e-02
```

2) Decision Tree

```
## n= 1021
##
## node), split, n, deviance, yval
##         * denotes terminal node
##
##  1) root 1021 161.4807000 12.02676
##    2) OverallQual< 6.5 633  51.0465400 11.81519
##      4) AllSF< 3314.5 293  18.3285300 11.64129
##        8) Neighborhood=BrDale,BrkSide,Edwards,IDOTRR,MeadowV,OldTown,SawyerW,SWISU 140   9.8231980 1
##         16) OverallQual< 3.5 10   0.7829115 11.01929 *
##         17) OverallQual>=3.5 130   6.5825690 11.53376 *
##        9) Neighborhood=Blueste,ClearCr,CollgCr,Crawfor,Gilbert,Mitchel,NAmes,NPkVill,Sawyer,Somerst,T
##      5) AllSF>=3314.5 340  16.2211300 11.96506
##       10) Neighborhood=BrDale,BrkSide,Edwards,IDOTRR,MeadowV,Mitchel,NAmes,NPkVill,OldTown,Sawyer,SWI
##       11) Neighborhood=ClearCr,CollgCr,Crawfor,Gilbert,NridgHt,NWAmes,SawyerW,Somerst,Timber,Veenker
##    3) OverallQual>=6.5 388  35.8779600 12.37191
##      6) AllSF< 4815.5 207   6.7910660 12.17948
##       12) AllSF< 3623 33   0.4690366 11.96968 *
##       13) AllSF>=3623 174   4.5939690 12.21927 *
##      7) AllSF>=4815.5 181  12.6557500 12.59198
##       14) Neighborhood=ClearCr,CollgCr,Crawfor,Edwards,Gilbert,Mitchel,NAmes,NWAmes,OldTown,SawyerW,S
##       15) Neighborhood=NoRidge,NridgHt,StoneBr,Veenker 82   5.1704670 12.74811
##         30) AllSF< 6494.5 62   1.8075070 12.65837 *
##         31) AllSF>=6494.5 20   1.3156410 13.02632 *
```



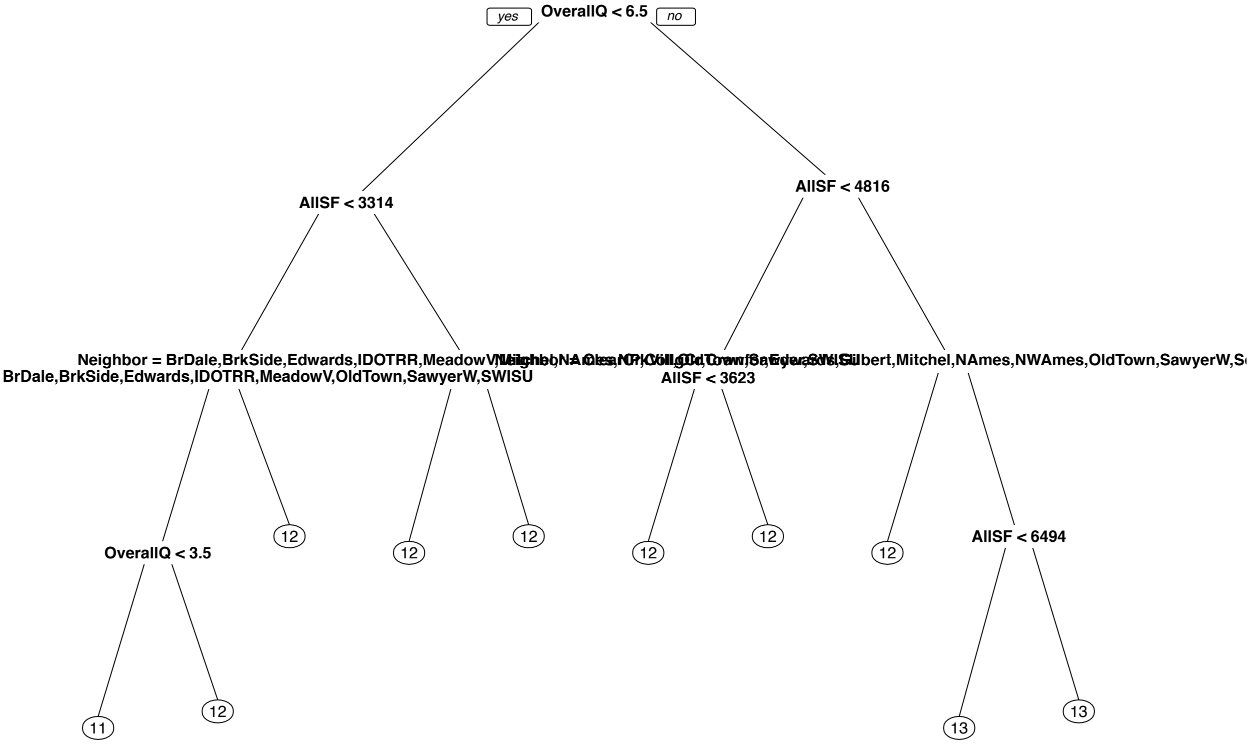


Calculate SSE and RMSE

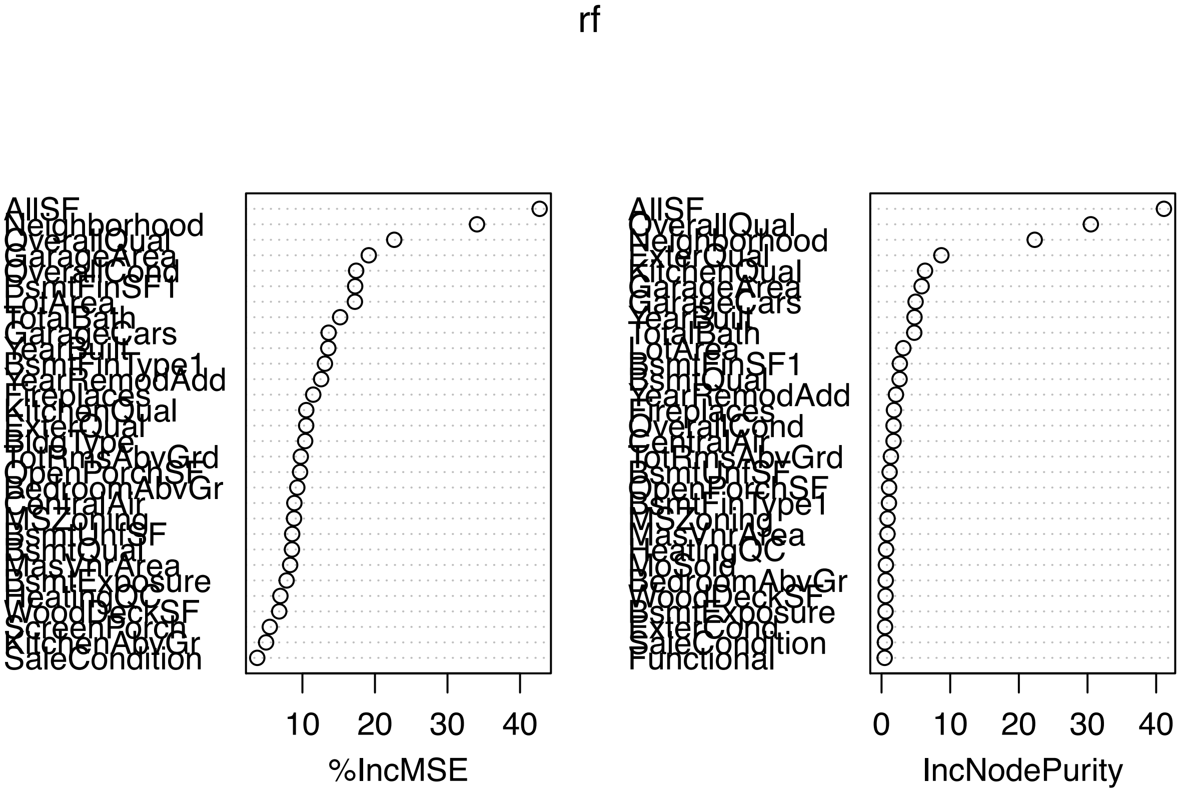
```
## [1] 18.77595
```

```
## [1] 0.2068086
```

Draw Tree using another way


yes
OverallQ < 6.5
no
AllSF < 3314
AllSF < 4816
Neighbor = BrDale,BrkSide,Edwards,IDOTRR,MeadowV,OldTown,SawyerW,SWISU
BrDale,BrkSide,Edwards,IDOTRR,MeadowV,OldTown,SawyerW,SWISU
Neighbor = BrDale,BrkSide,ClearCr,Edwards,Gilbert,Mitchel,NAmes,NWAmes,OldTown,SawyerW,S
AllSF < 3623
OverallQ < 3.5
AllSF < 6494
12
12
12
12
12
12
11
12
13
13

3) Random Forest



Strong Feature shown in Random Forest are:

```
##  [1] "AllSF"         "OverallQual"   "Neighborhood"  "GarageArea"
##  [5] "YearBuilt"     "ExterQual"     "TotalBath"     "KitchenQual"  
##  [9] "GarageCars"    "YearRemodAdd"  "LotArea"       "BsmtFinSF1"   
## [13] "BsmtQual"      "OverallCond"   "Fireplaces"    "BsmtFinType1" 
## [17] "TotRmsAbvGrd"  "CentralAir"    "MSZoning"      "OpenPorchSF"  
## [21] "BsmtUnfSF"     "BedroomAbvGr"  "HeatingQC"     "MasVnrArea"   
## [25] "BldgType"      "BsmtExposure"  "WoodDeckSF"    "SaleCondition"
## [29] "KitchenAbvGr"  "ScreenPorch"   "LotConfig"     "Functional"   
## [33] "Condition1"    "RoofStyle"     "BsmtFinType2"  "LandContour"  
## [37] "SaleType"      "BsmtFinSF2"    "ExterCond"     "Condition2"   
## [41] "X3SsnPorch"    "LandSlope"     "Street"        "Utilities"    
## [45] "YrSold"        "RoofMatl"      "LowQualFinSF"  "PoolArea"     
## [49] "MoSold"        "Heating"
```

Calculate SSE and RMSE ==> Find out random forest has a obviously decrese on SSE and RMSE

```
## [1] 8.467138
```

```
## [1] 0.1388788
```

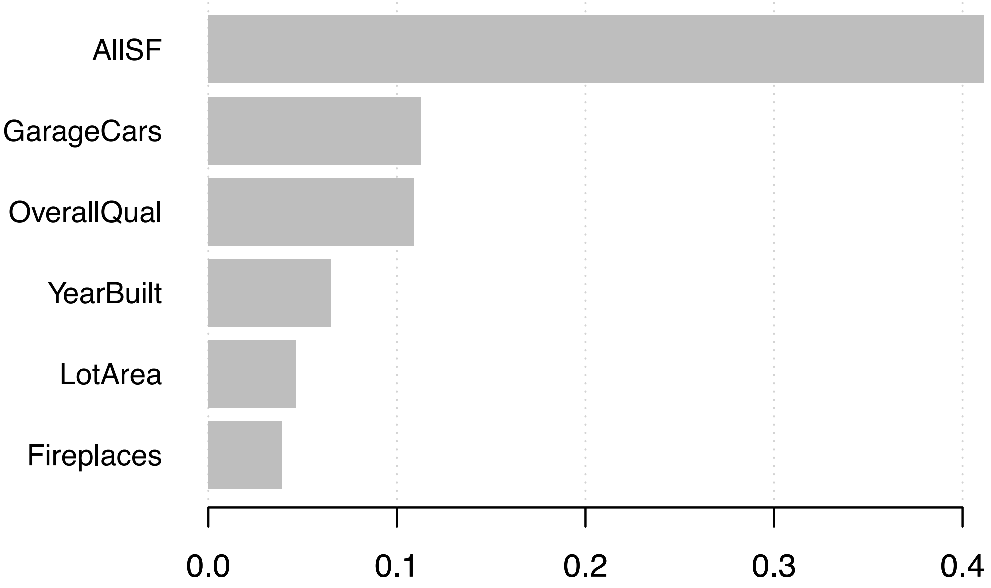
4) Boosting

```
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 0 extra nodes, 0
## [1]  train-rmse:8.081903 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 2 extra nodes, 0
## [2]  train-rmse:5.665978 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 2 extra nodes, 0
## [3]  train-rmse:3.974063 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 2 extra nodes, 0
## [4]  train-rmse:2.790202 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 extra nodes, 0
## [5]  train-rmse:1.961632 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8 extra nodes, 0
## [6]  train-rmse:1.381762 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 14 extra nodes, 0
## [7]  train-rmse:0.976381 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 28 extra nodes, 0
## [8]  train-rmse:0.692852 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 48 extra nodes, 0
## [9]  train-rmse:0.495095 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 74 extra nodes, 0
## [10] train-rmse:0.356784 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 118 extra nodes,
## [11] train-rmse:0.260117 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 96 extra nodes, 0
## [12] train-rmse:0.192877 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 154 extra nodes,
## [13] train-rmse:0.146042 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 114 extra nodes,
## [14] train-rmse:0.113968 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 120 extra nodes,
## [15] train-rmse:0.091927 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 142 extra nodes,
## [16] train-rmse:0.076759 
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 156 extra nodes,
## [17] train-rmse:0.066555
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 198 extra nodes,
## [18] train-rmse:0.057948
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 174 extra nodes,
## [19] train-rmse:0.052218
## [21:46:03] amalgamation/../src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 128 extra nodes,
## [20] train-rmse:0.048780
```

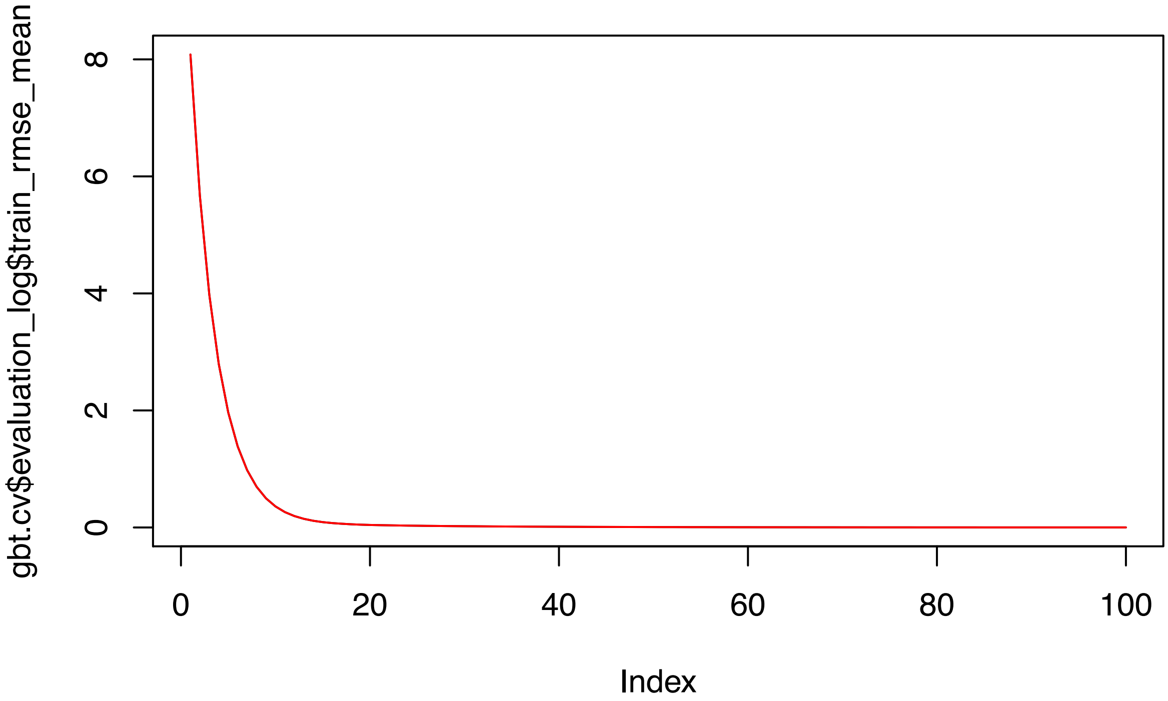
Show first sixth important features and plot

```
##        Feature       Gain      Cover   Frequency
## 1:       AllSF 0.41133966 0.18331224 0.077020202
## 2:  GarageCars 0.11276111 0.01282657 0.006313131
## 3: OverallQual 0.10917371 0.06924268 0.036616162
## 4:   YearBuilt 0.06504738 0.06723086 0.050505051
## 5:     LotArea 0.04637810 0.09753103 0.098484848
## 6:  Fireplaces 0.03920406 0.01602849 0.011363636
```

```
## Warning: package 'Ckmeans.1d.dp' was built under R version 3.3.2
```



Using Cross Validation to find nround best used in xgboost funtion to minimum prediction error



Calculate SSE and RMSE

```
## [1] 9.259196
```

```
## [1] 0.1452293
```