Model

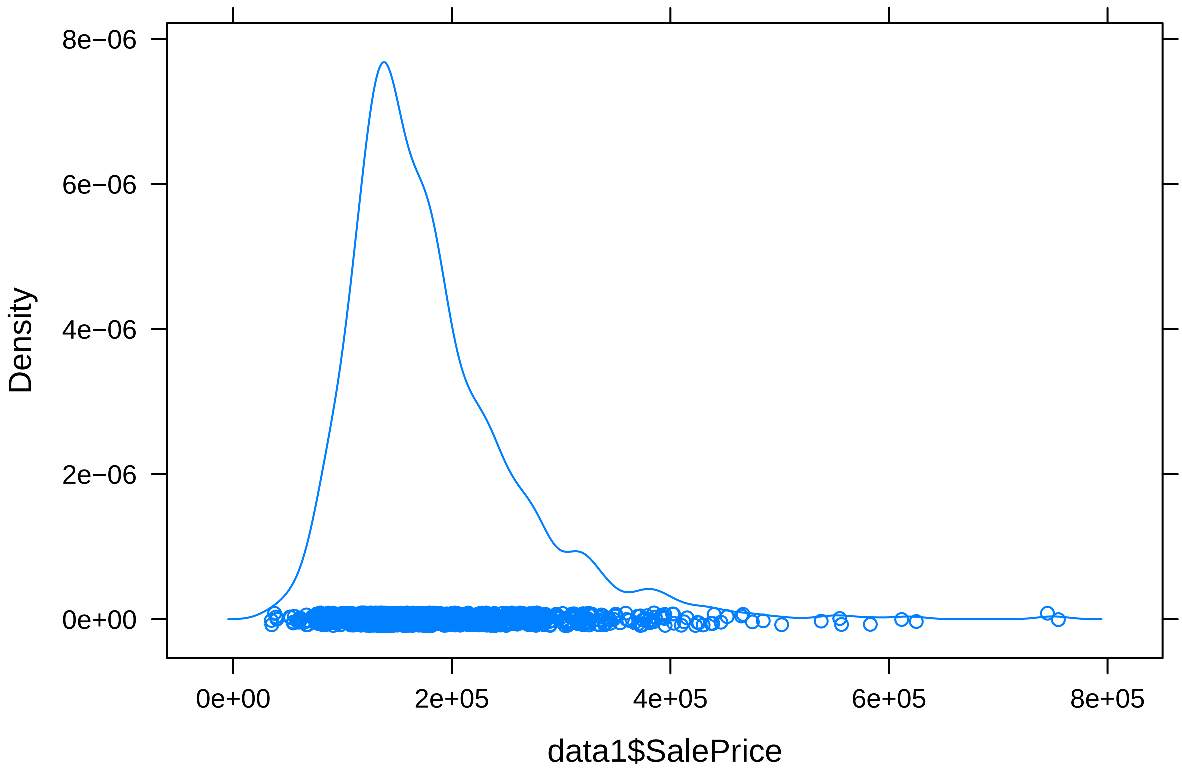
Jz

```r
data1 <- read.csv("train.csv", header = TRUE, na.strings = "NA")
data2 <- read.csv("test.csv", header = TRUE, na.strings = "NA")
data1 <- data1[, -1] #exclude id column
dim(data1) #1460 rows, 80 variables
```

```
## [1] 1460   80
```

```r
set.seed(11)
```

```r
library("lattice")
densityplot(data1$SalePrice)
```




```r
#it seems like a normal distribution
```

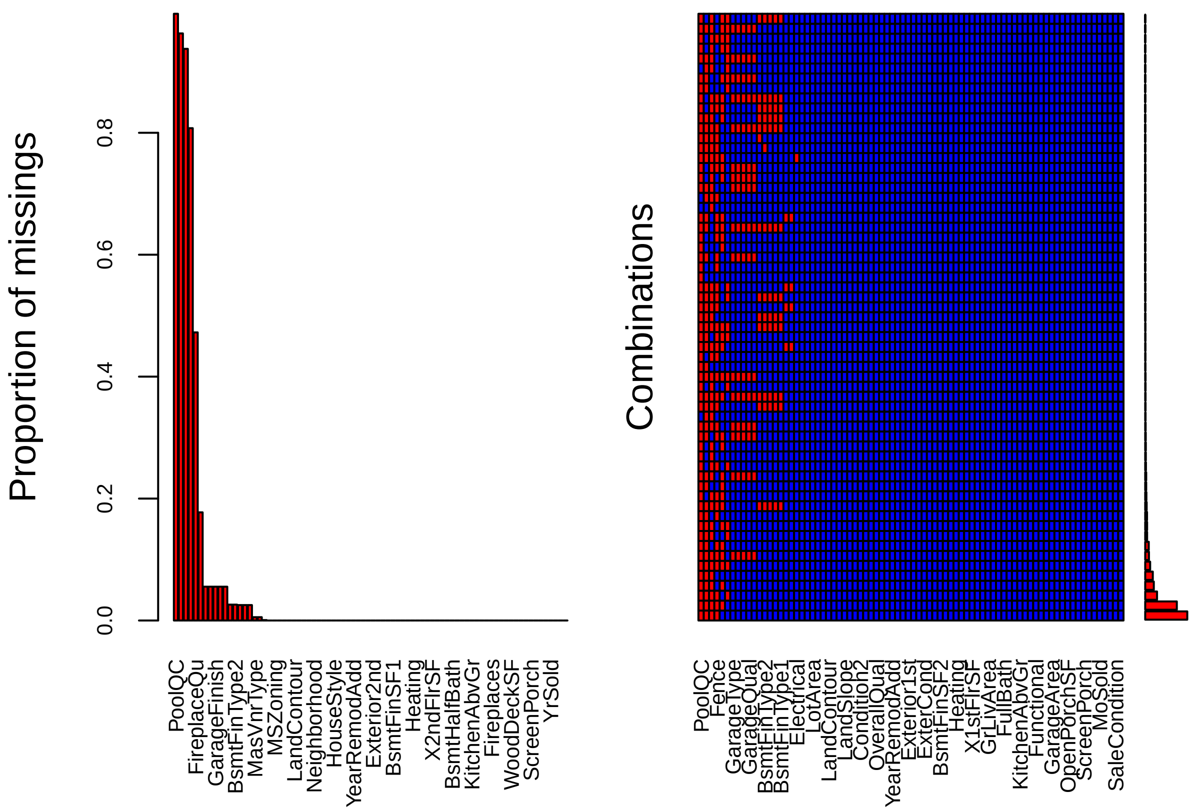
```r
#understand proportion of missing data
missing <- function(x){
sum(is.na(x))/length(x)
}
sort(sapply(data1, missing), decreasing = TRUE)[1:10]
```

```
##       PoolQC  MiscFeature        Alley        Fence  FireplaceQu
##   0.99520548   0.96301370   0.93767123   0.80753425   0.47260274
##  LotFrontage   GarageType  GarageYrBlt GarageFinish   GarageQual
##   0.17739726   0.05547945   0.05547945   0.05547945   0.05547945
```

```r
#visualize missing data
library(VIM)
```

```
## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

## Warning: package 'data.table' was built under R version 3.4.2

## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.

## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep
```

```r
missing_plot <- aggr(data1, col = c("blue", "red"),
sortVars = TRUE, labels = names(data1),
cex.axis = 0.7, gap = 3)
```




```
##
##  Variables sorted by number of missings:
##       Variable        Count
##         PoolQC 0.9952054795
##    MiscFeature 0.9630136986
```

```
##          Alley 0.9376712329
##          Fence 0.8075342466
##    FireplaceQu 0.4726027397
##    LotFrontage 0.1773972603
##     GarageType 0.0554794521
##    GarageYrBlt 0.0554794521
##   GarageFinish 0.0554794521
##     GarageQual 0.0554794521
##     GarageCond 0.0554794521
##   BsmtExposure 0.0260273973
##   BsmtFinType2 0.0260273973
##       BsmtQual 0.0253424658
##       BsmtCond 0.0253424658
##   BsmtFinType1 0.0253424658
##     MasVnrType 0.0054794521
##     MasVnrArea 0.0054794521
##     Electrical 0.0006849315
##     MSSubClass 0.0000000000
##       MSZoning 0.0000000000
##        LotArea 0.0000000000
##         Street 0.0000000000
##       LotShape 0.0000000000
##    LandContour 0.0000000000
##      Utilities 0.0000000000
##      LotConfig 0.0000000000
##      LandSlope 0.0000000000
##   Neighborhood 0.0000000000
##     Condition1 0.0000000000
##     Condition2 0.0000000000
##       BldgType 0.0000000000
##     HouseStyle 0.0000000000
##    OverallQual 0.0000000000
##    OverallCond 0.0000000000
##      YearBuilt 0.0000000000
##   YearRemodAdd 0.0000000000
##      RoofStyle 0.0000000000
##       RoofMatl 0.0000000000
##    Exterior1st 0.0000000000
##    Exterior2nd 0.0000000000
##      ExterQual 0.0000000000
##      ExterCond 0.0000000000
##     Foundation 0.0000000000
##     BsmtFinSF1 0.0000000000
##     BsmtFinSF2 0.0000000000
##      BsmtUnfSF 0.0000000000
##    TotalBsmtSF 0.0000000000
##        Heating 0.0000000000
##      HeatingQC 0.0000000000
##     CentralAir 0.0000000000
##      X1stFlrSF 0.0000000000
##      X2ndFlrSF 0.0000000000
##   LowQualFinSF 0.0000000000
##      GrLivArea 0.0000000000
##   BsmtFullBath 0.0000000000
```

```
##   BsmtHalfBath 0.0000000000
##       FullBath 0.0000000000
##       HalfBath 0.0000000000
##   BedroomAbvGr 0.0000000000
##   KitchenAbvGr 0.0000000000
##    KitchenQual 0.0000000000
##   TotRmsAbvGrd 0.0000000000
##     Functional 0.0000000000
##     Fireplaces 0.0000000000
##     GarageCars 0.0000000000
##     GarageArea 0.0000000000
##     PavedDrive 0.0000000000
##     WoodDeckSF 0.0000000000
##    OpenPorchSF 0.0000000000
##  EnclosedPorch 0.0000000000
##     X3SsnPorch 0.0000000000
##    ScreenPorch 0.0000000000
##       PoolArea 0.0000000000
##        MiscVal 0.0000000000
##         MoSold 0.0000000000
##         YrSold 0.0000000000
##       SaleType 0.0000000000
##  SaleCondition 0.0000000000
##      SalePrice 0.0000000000
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
data1 <- select(data1, -c(PoolQC, MiscFeature, Alley, Fence,
FireplaceQu, LotFrontage))
library(mice)
```

```
## Warning: package 'mice' was built under R version 3.4.2
```

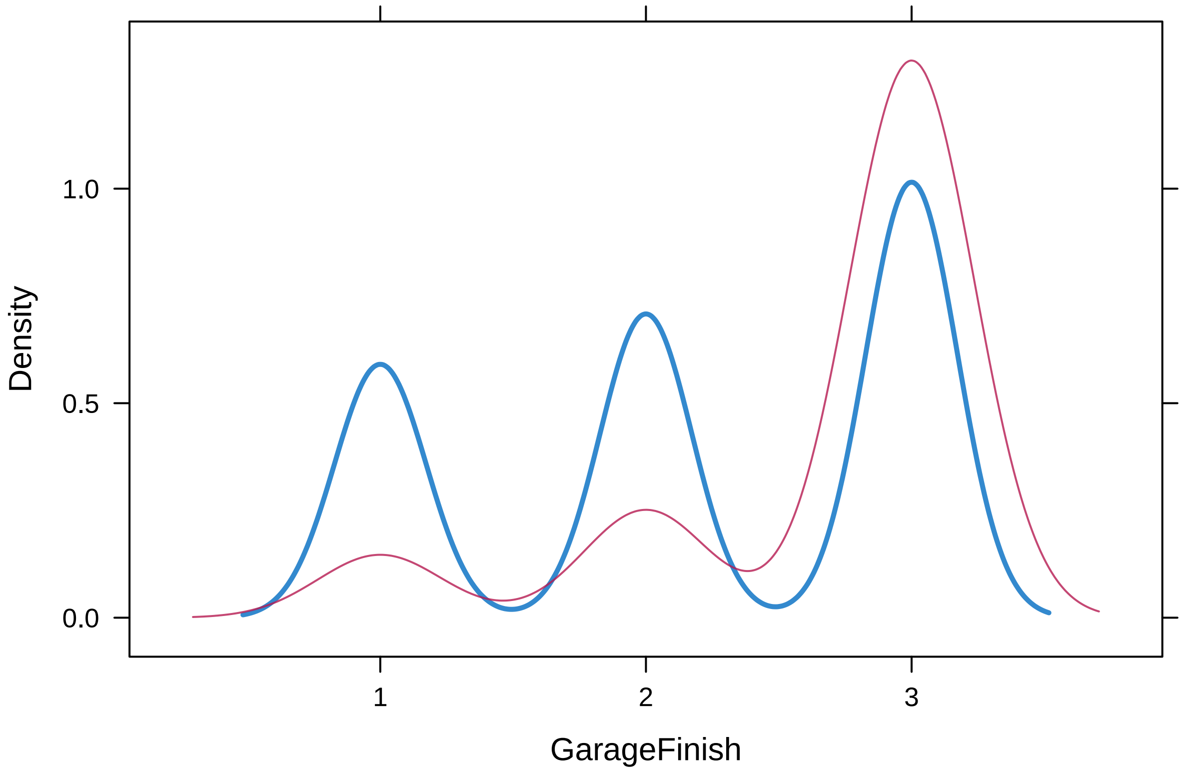
```r
#using cart
imp_data <- mice(data1, m = 1, method = "cart", printFlag = FALSE)
#because of large numbers of unbalanced factor variables, when they change to dummy vairables, there is
table(imp_data$imp$ GarageFinish)
```

```
##
## Fin RFn Unf
##   7  12  62
```

```r
table(data1$ GarageFinish)
```

```
##
## Fin RFn Unf
## 352 422 605
```

```r
densityplot(imp_data, ~ GarageFinish) #from pattern it is acceptable
```




```r
full_data1 <- complete(imp_data)
# then double check no missing data
sort(sapply(full_data1, missing), decreasing = TRUE)[1:5] #no missing data
```

```
## MSSubClass   MSZoning    LotArea     Street   LotShape
##          0          0          0          0          0
```

```r
set.seed(11)
train <- sample(1:nrow(full_data1), nrow(full_data1)/10*6)
test <- -train
traindata <- full_data1[train, ]
testdata <- full_data1[test, ]
ols_model <- lm(SalePrice ~., data = traindata)
```

```
## Warning: contrasts dropped from factor Condition2 due to missing levels

## Warning: contrasts dropped from factor RoofMatl due to missing levels

## Warning: contrasts dropped from factor Exterior1st due to missing levels

## Warning: contrasts dropped from factor Exterior2nd due to missing levels
```

```r
summary(ols_model)
```

```
##
```

```
## Call:
## lm(formula = SalePrice ~ ., data = traindata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -132938   -9996     431   10373  136429
##
## Coefficients: (4 not defined because of singularities)
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.001e+05  1.365e+06   0.513 0.608110
## MSSubClass          3.817e+01  1.238e+02   0.308 0.757876
## MSZoning2           2.711e+04  1.788e+04   1.516 0.129999
## MSZoning3           9.660e+03  1.923e+04   0.502 0.615618
## MSZoning4           2.074e+04  1.613e+04   1.286 0.198985
## MSZoning5           1.870e+04  1.550e+04   1.206 0.228141
## LotArea             1.172e+00  2.577e-01   4.547 6.48e-06 ***
## Street2             1.448e+04  1.532e+04   0.945 0.344821
## LotShape2          -2.191e+03  5.729e+03  -0.382 0.702252
## LotShape3          -1.344e+02  1.272e+04  -0.011 0.991579
## LotShape4          -1.551e+03  2.102e+03  -0.738 0.460921
## LandContour2        2.919e+03  6.940e+03   0.421 0.674200
## LandContour3       -1.619e+04  8.868e+03  -1.826 0.068364 .
## LandContour4        2.552e+03  5.013e+03   0.509 0.610909
## Utilities2         -1.959e+04  2.860e+04  -0.685 0.493597
## LotConfig2          8.949e+03  4.099e+03   2.183 0.029367 *
## LotConfig3         -5.447e+03  5.162e+03  -1.055 0.291746
## LotConfig4         -8.308e+03  1.374e+04  -0.605 0.545492
## LotConfig5         -1.845e+02  2.273e+03  -0.081 0.935334
## LandSlope2          5.435e+02  5.395e+03   0.101 0.919781
## LandSlope3         -3.631e+04  1.362e+04  -2.665 0.007880 **
## Neighborhood2       9.371e+03  2.907e+04   0.322 0.747279
## Neighborhood3      -8.370e+03  1.462e+04  -0.573 0.567160
## Neighborhood4      -3.232e+02  1.342e+04  -0.024 0.980792
## Neighborhood5      -1.763e+04  1.276e+04  -1.382 0.167505
## Neighborhood6      -1.003e+04  9.998e+03  -1.003 0.316134
## Neighborhood7       1.430e+04  1.183e+04   1.208 0.227335
## Neighborhood8      -1.544e+04  1.113e+04  -1.387 0.166028
## Neighborhood9      -1.266e+04  1.048e+04  -1.208 0.227561
## Neighborhood10     -7.921e+03  1.450e+04  -0.546 0.584972
## Neighborhood11     -2.030e+03  1.659e+04  -0.122 0.902617
## Neighborhood12     -2.331e+04  1.134e+04  -2.055 0.040289 *
## Neighborhood13     -1.419e+04  1.089e+04  -1.303 0.192934
## Neighborhood14      2.346e+04  1.108e+04   2.118 0.034587 *
## Neighborhood15      6.082e+03  2.488e+04   0.245 0.806917
## Neighborhood16      3.028e+03  1.018e+04   0.297 0.766278
## Neighborhood17     -2.228e+04  1.092e+04  -2.041 0.041657 *
## Neighborhood18     -1.023e+04  1.322e+04  -0.774 0.439136
## Neighborhood19     -1.369e+04  1.125e+04  -1.217 0.224072
## Neighborhood20     -1.645e+03  1.058e+04  -0.156 0.876457
## Neighborhood21     -6.167e+03  1.178e+04  -0.524 0.600743
## Neighborhood22      2.686e+04  1.140e+04   2.356 0.018774 *
## Neighborhood23     -1.114e+04  1.323e+04  -0.842 0.400190
## Neighborhood24     -1.314e+04  1.164e+04  -1.128 0.259575
## Neighborhood25     -8.519e+03  1.476e+04  -0.577 0.563925
```

```
## Condition12          6.321e+03  6.698e+03   0.944 0.345630
## Condition13          1.097e+04  5.637e+03   1.947 0.051979 .
## Condition14          1.618e+04  1.392e+04   1.162 0.245509
## Condition15          7.447e+03  9.735e+03   0.765 0.444573
## Condition16         -2.338e+04  1.423e+04  -1.642 0.100994
## Condition17          1.832e+03  1.045e+04   0.175 0.860843
## Condition18          2.439e+02  2.371e+04   0.010 0.991797
## Condition19         -7.858e+03  1.708e+04  -0.460 0.645621
## Condition2Feedr     -1.231e+04  3.753e+04  -0.328 0.742979
## Condition2Norm       6.775e+03  3.132e+04   0.216 0.828809
## Condition2PosN      -4.619e+05  4.188e+04 -11.030  < 2e-16 ***
## Condition2RRAe      -1.339e+04  6.892e+04  -0.194 0.845978
## Condition2RRAn      -4.284e+03  4.001e+04  -0.107 0.914748
## Condition2RRNn       1.802e+04  3.700e+04   0.487 0.626319
## BldgType2           -9.924e+03  1.902e+04  -0.522 0.602003
## BldgType3           -1.327e+04  1.014e+04  -1.309 0.191064
## BldgType4           -2.225e+04  1.438e+04  -1.547 0.122277
## BldgType5           -1.956e+04  1.307e+04  -1.497 0.134841
## HouseStyle2          1.832e+04  9.916e+03   1.848 0.065102 .
## HouseStyle3          1.413e+04  6.013e+03   2.349 0.019096 *
## HouseStyle4         -1.108e+04  1.414e+04  -0.784 0.433427
## HouseStyle5          2.180e+03  1.547e+04   0.141 0.888001
## HouseStyle6         -4.180e+03  4.804e+03  -0.870 0.384579
## HouseStyle7          6.889e+03  8.686e+03   0.793 0.427998
## HouseStyle8          8.148e+03  7.372e+03   1.105 0.269424
## OverallQual          6.723e+03  1.360e+03   4.945 9.71e-07 ***
## OverallCond          6.314e+03  1.187e+03   5.318 1.45e-07 ***
## YearBuilt            4.189e+02  1.092e+02   3.838 0.000136 ***
## YearRemodAdd         3.962e+01  7.438e+01   0.533 0.594502
## RoofStyle2           2.930e+03  2.095e+04   0.140 0.888856
## RoofStyle3           7.390e+03  3.154e+04   0.234 0.814832
## RoofStyle4           1.805e+03  2.108e+04   0.086 0.931804
## RoofStyle5           1.625e+04  2.614e+04   0.622 0.534483
## RoofStyle6                  NA         NA      NA       NA
## RoofMatlMembran      9.508e+04  3.522e+04   2.700 0.007123 **
## RoofMatlMetal        5.745e+04  3.279e+04   1.752 0.080177 .
## RoofMatlRoll        -2.160e+04  2.746e+04  -0.787 0.431804
## RoofMatlTar&Grv     -5.253e+03  1.971e+04  -0.267 0.789884
## RoofMatlWdShake     -1.347e+04  2.437e+04  -0.553 0.580594
## RoofMatlWdShngl      5.053e+04  1.262e+04   4.003 6.97e-05 ***
## Exterior1stBrkComm -3.430e+04  4.029e+04  -0.851 0.394912
## Exterior1stBrkFace -1.924e+04  1.987e+04  -0.968 0.333250
## Exterior1stCemntBd -1.170e+04  1.265e+04  -0.924 0.355691
## Exterior1stHdBoard -3.516e+04  2.004e+04  -1.754 0.079886 .
## Exterior1stImStucc -7.667e+04  3.158e+04  -2.428 0.015468 *
## Exterior1stMetalSd -2.667e+04  2.207e+04  -1.209 0.227213
## Exterior1stPlywood -3.994e+04  1.983e+04  -2.014 0.044419 *
## Exterior1stStone   -3.207e+04  3.107e+04  -1.032 0.302357
## Exterior1stStucco  -1.720e+04  2.152e+04  -0.799 0.424429
## Exterior1stVinylSd -3.527e+04  2.073e+04  -1.701 0.089335 .
## Exterior1stWd Sdng -3.730e+04  1.921e+04  -1.941 0.052663 .
## Exterior1stWdShing -3.348e+04  2.065e+04  -1.622 0.105349
## Exterior2ndAsphShn  2.557e+04  3.459e+04   0.739 0.459965
## Exterior2ndBrk Cmn  3.245e+04  3.476e+04   0.933 0.350949
```

```
## Exterior2ndBrkFace  2.353e+04  2.130e+04   1.105 0.269696
## Exterior2ndCmentBd         NA         NA      NA       NA
## Exterior2ndHdBoard  2.844e+04  2.032e+04   1.400 0.162112
## Exterior2ndImStucc  5.147e+04  2.208e+04   2.332 0.020023 *
## Exterior2ndMetalSd  1.886e+04  2.238e+04   0.843 0.399673
## Exterior2ndOther    1.272e+04  3.213e+04   0.396 0.692233
## Exterior2ndPlywood  2.665e+04  1.976e+04   1.349 0.177856
## Exterior2ndStone    1.824e+04  2.526e+04   0.722 0.470581
## Exterior2ndStucco   1.167e+04  2.188e+04   0.533 0.593984
## Exterior2ndVinylSd  2.944e+04  2.105e+04   1.399 0.162382
## Exterior2ndWd Sdng  2.932e+04  1.956e+04   1.499 0.134274
## Exterior2ndWd Shng  2.581e+04  2.019e+04   1.278 0.201538
## MasVnrType2        -6.692e+03  1.194e+04  -0.560 0.575354
## MasVnrType3         8.045e+01  1.187e+04   0.007 0.994594
## MasVnrType4         3.990e+03  1.226e+04   0.325 0.745040
## MasVnrArea          2.205e+01  7.667e+00   2.877 0.004152 **
## ExterQual2         -1.272e+03  1.648e+04  -0.077 0.938473
## ExterQual3         -1.892e+04  6.436e+03  -2.940 0.003396 **
## ExterQual4         -1.848e+04  7.079e+03  -2.611 0.009227 **
## ExterCond2          1.862e+03  3.474e+04   0.054 0.957279
## ExterCond3         -5.321e+03  3.411e+04  -0.156 0.876102
## ExterCond4          2.126e+04  4.393e+04   0.484 0.628554
## ExterCond5         -1.607e+03  3.433e+04  -0.047 0.962678
## Foundation2         1.629e+03  4.386e+03   0.371 0.710401
## Foundation3         1.797e+03  4.721e+03   0.381 0.703532
## Foundation4         1.320e+04  1.053e+04   1.254 0.210205
## Foundation5         2.202e+04  1.709e+04   1.289 0.197928
## Foundation6        -4.899e+04  2.632e+04  -1.861 0.063134 .
## BsmtQual2          -1.923e+04  8.613e+03  -2.233 0.025901 *
## BsmtQual3          -2.482e+04  4.538e+03  -5.468 6.47e-08 ***
## BsmtQual4          -2.196e+04  5.629e+03  -3.902 0.000105 ***
## BsmtCond2          -1.557e+03  7.226e+03  -0.215 0.829451
## BsmtCond3           7.419e+04  3.520e+04   2.108 0.035454 *
## BsmtCond4           3.969e+03  5.772e+03   0.688 0.491901
## BsmtExposure2       1.501e+04  4.111e+03   3.650 0.000283 ***
## BsmtExposure3       1.261e+03  3.857e+03   0.327 0.743898
## BsmtExposure4      -1.763e+03  2.777e+03  -0.635 0.525771
## BsmtFinType12       1.627e+03  3.569e+03   0.456 0.648547
## BsmtFinType13       6.241e+03  3.200e+03   1.950 0.051555 .
## BsmtFinType14      -4.160e+03  4.940e+03  -0.842 0.399937
## BsmtFinType15      -3.378e+03  3.897e+03  -0.867 0.386345
## BsmtFinType16      -3.212e+02  3.818e+03  -0.084 0.932982
## BsmtFinSF1          4.089e+01  6.370e+00   6.418 2.65e-10 ***
## BsmtFinType22      -2.051e+04  9.893e+03  -2.073 0.038539 *
## BsmtFinType23      -1.279e+04  1.259e+04  -1.016 0.310063
## BsmtFinType24      -2.034e+04  9.625e+03  -2.113 0.034959 *
## BsmtFinType25      -2.001e+04  8.997e+03  -2.224 0.026508 *
## BsmtFinType26      -1.231e+04  9.643e+03  -1.277 0.202117
## BsmtFinSF2          3.210e+01  1.125e+01   2.852 0.004477 **
## BsmtUnfSF           2.146e+01  5.845e+00   3.671 0.000261 ***
## TotalBsmtSF                NA         NA      NA       NA
## Heating2            8.038e+03  2.573e+04   0.312 0.754806
## Heating3            6.932e+03  2.778e+04   0.250 0.803044
## Heating4            1.090e+04  2.994e+04   0.364 0.715908
```

```
## Heating5          -6.715e+03  3.714e+04  -0.181 0.856581
## Heating6           1.890e+04  3.431e+04   0.551 0.581897
## HeatingQC2        -3.306e+03  6.140e+03  -0.538 0.590511
## HeatingQC3        -3.521e+03  2.724e+03  -1.292 0.196685
## HeatingQC4         8.219e+03  2.992e+04   0.275 0.783624
## HeatingQC5        -3.401e+03  2.787e+03  -1.220 0.222751
## CentralAir2        3.340e+03  5.626e+03   0.594 0.552951
## Electrical2        4.050e+03  8.270e+03   0.490 0.624471
## Electrical3       -2.382e+04  3.406e+04  -0.699 0.484576
## Electrical4       -5.293e+04  5.377e+04  -0.984 0.325301
## Electrical5        3.053e+02  4.080e+03   0.075 0.940377
## X1stFlrSF          5.606e+01  7.232e+00   7.751 3.50e-14 ***
## X2ndFlrSF          7.494e+01  7.090e+00  10.569  < 2e-16 ***
## LowQualFinSF      -4.300e-01  2.346e+01  -0.018 0.985381
## GrLivArea                 NA         NA      NA       NA
## BsmtFullBath      -1.957e+03  2.526e+03  -0.775 0.438784
## BsmtHalfBath      -2.318e+03  3.872e+03  -0.599 0.549650
## FullBath           2.740e+03  3.161e+03   0.867 0.386364
## HalfBath           2.971e+03  2.726e+03   1.090 0.276140
## BedroomAbvGr      -3.321e+03  1.822e+03  -1.822 0.068892 .
## KitchenAbvGr      -1.539e+04  8.383e+03  -1.836 0.066845 .
## KitchenQual2      -2.653e+04  7.975e+03  -3.326 0.000930 ***
## KitchenQual3      -3.063e+04  4.572e+03  -6.700 4.52e-11 ***
## KitchenQual4      -2.738e+04  4.989e+03  -5.487 5.84e-08 ***
## TotRmsAbvGrd      -6.632e+01  1.274e+03  -0.052 0.958494
## Functional2        1.228e+04  2.262e+04   0.543 0.587343
## Functional3        1.269e+04  1.193e+04   1.064 0.287934
## Functional4        1.686e+04  1.205e+04   1.400 0.162063
## Functional5       -6.685e+03  1.470e+04  -0.455 0.649488
## Functional6       -6.227e+04  3.629e+04  -1.716 0.086642 .
## Functional7        2.184e+04  1.052e+04   2.075 0.038383 *
## Fireplaces         7.538e+02  1.742e+03   0.433 0.665432
## GarageType2        8.973e+03  1.204e+04   0.745 0.456474
## GarageType3        1.733e+04  1.417e+04   1.223 0.221768
## GarageType4        1.294e+04  1.272e+04   1.017 0.309294
## GarageType5        2.700e+04  1.812e+04   1.490 0.136729
## GarageType6        1.613e+04  1.205e+04   1.338 0.181375
## GarageYrBlt        2.670e+01  8.352e+01   0.320 0.749355
## GarageFinish2     -3.082e+02  2.577e+03  -0.120 0.904832
## GarageFinish3      1.521e+03  3.262e+03   0.466 0.641232
## GarageCars         5.112e+03  2.904e+03   1.760 0.078807 .
## GarageArea         2.123e+00  1.069e+01   0.199 0.842587
## GarageQual2       -7.952e+04  2.149e+04  -3.700 0.000234 ***
## GarageQual3       -6.430e+04  2.424e+04  -2.653 0.008180 **
## GarageQual4       -8.597e+04  3.220e+04  -2.669 0.007788 **
## GarageQual5       -7.285e+04  2.188e+04  -3.330 0.000918 ***
## GarageCond2        6.534e+04  2.873e+04   2.274 0.023272 *
## GarageCond3        4.655e+04  3.068e+04   1.517 0.129665
## GarageCond4        5.626e+04  2.816e+04   1.998 0.046120 *
## GarageCond5        6.438e+04  2.816e+04   2.286 0.022589 *
## PavedDrive2        1.585e+03  7.683e+03   0.206 0.836598
## PavedDrive3        1.318e+03  5.408e+03   0.244 0.807509
## WoodDeckSF         1.277e+01  7.961e+00   1.604 0.109149
## OpenPorchSF        1.304e+01  1.529e+01   0.853 0.393970
```

```
## EnclosedPorch      -3.263e-01  1.615e+01  -0.020 0.983883
## X3SsnPorch          1.862e+01  2.868e+01   0.649 0.516288
## ScreenPorch         1.093e+01  1.674e+01   0.653 0.513739
## PoolArea            9.252e+01  2.263e+01   4.089 4.88e-05 ***
## MiscVal            -1.262e+00  5.772e+00  -0.219 0.827053
## MoSold             -2.829e+02  3.209e+02  -0.882 0.378295
## YrSold             -8.443e+02  6.684e+02  -1.263 0.206987
## SaleType2           5.012e+04  2.654e+04   1.888 0.059423 .
## SaleType3           6.261e+03  1.144e+04   0.547 0.584228
## SaleType4           6.678e+03  1.211e+04   0.552 0.581366
## SaleType5          -3.562e+03  1.648e+04  -0.216 0.828931
## SaleType6           1.218e+04  1.585e+04   0.769 0.442456
## SaleType7           2.347e+04  1.647e+04   1.425 0.154571
## SaleType8           7.415e+03  1.739e+04   0.426 0.669977
## SaleType9           1.552e+03  5.393e+03   0.288 0.773641
## SaleCondition2      3.189e+04  3.121e+04   1.022 0.307217
## SaleCondition3      1.161e+03  1.322e+04   0.088 0.930070
## SaleCondition4     -2.420e+03  8.385e+03  -0.289 0.773004
## SaleCondition5      6.824e+03  3.700e+03   1.845 0.065555 .
## SaleCondition6      7.737e+02  1.546e+04   0.050 0.960109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21980 on 653 degrees of freedom
## Multiple R-squared:  0.9432, Adjusted R-squared:  0.9238
## F-statistic: 48.82 on 222 and 653 DF,  p-value: < 2.2e-16
```

```r
# adjusted R^2 = 92.6%, not bad; some vairables are have too big p-value
ols_model_rmse <- sqrt(mean(ols_model$residuals ^2))
ols_model_rmse #18975.6
```

```
## [1] 18975.6
```

```r
ols_model2 <- lm(SalePrice ~ LotArea + OverallQual + OverallCond
+ YearBuilt + MasVnrArea + BsmtQual +BsmtFinSF1 +
BsmtFinSF2 +BsmtUnfSF + X1stFlrSF + X2ndFlrSF +
KitchenQual + KitchenAbvGr +BedroomAbvGr+
GarageCars +PoolArea,
data = traindata)
summary(ols_model2)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + OverallQual + OverallCond +
##     YearBuilt + MasVnrArea + BsmtQual + BsmtFinSF1 + BsmtFinSF2 +
##     BsmtUnfSF + X1stFlrSF + X2ndFlrSF + KitchenQual + KitchenAbvGr +
##     BedroomAbvGr + GarageCars + PoolArea, data = traindata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -429522  -12437     928   12875  201334
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.446e+05  1.262e+05  -5.107 4.04e-07 ***
```

```
## LotArea       9.655e-01  2.351e-01   4.106 4.41e-05 ***
## OverallQual   1.116e+04  1.419e+03   7.862 1.14e-14 ***
## OverallCond   6.849e+03  1.072e+03   6.387 2.78e-10 ***
## YearBuilt     3.436e+02  6.247e+01   5.500 5.01e-08 ***
## MasVnrArea    1.971e+01  6.863e+00   2.872 0.004182 **
## BsmtQual2    -3.976e+04  9.120e+03  -4.360 1.46e-05 ***
## BsmtQual3    -4.039e+04  5.098e+03  -7.923 7.23e-15 ***
## BsmtQual4    -4.474e+04  6.119e+03  -7.312 6.06e-13 ***
## BsmtFinSF1    3.791e+01  5.008e+00   7.572 9.56e-14 ***
## BsmtFinSF2    2.262e+01  7.662e+00   2.952 0.003241 **
## BsmtUnfSF     2.065e+01  4.871e+00   4.238 2.50e-05 ***
## X1stFlrSF     5.819e+01  5.482e+00  10.615  < 2e-16 ***
## X2ndFlrSF     5.927e+01  3.818e+00  15.523  < 2e-16 ***
## KitchenQual2 -2.912e+04  8.537e+03  -3.411 0.000677 ***
## KitchenQual3 -2.965e+04  5.229e+03  -5.670 1.95e-08 ***
## KitchenQual4 -3.559e+04  5.714e+03  -6.228 7.39e-10 ***
## KitchenAbvGr -2.182e+04  5.225e+03  -4.176 3.28e-05 ***
## BedroomAbvGr -4.296e+03  1.710e+03  -2.512 0.012179 *
## GarageCars    8.942e+03  2.002e+03   4.468 8.97e-06 ***
## PoolArea      7.349e+01  2.368e+01   3.104 0.001971 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30890 on 855 degrees of freedom
## Multiple R-squared:  0.853,  Adjusted R-squared:  0.8496
## F-statistic: 248.1 on 20 and 855 DF,  p-value: < 2.2e-16
```

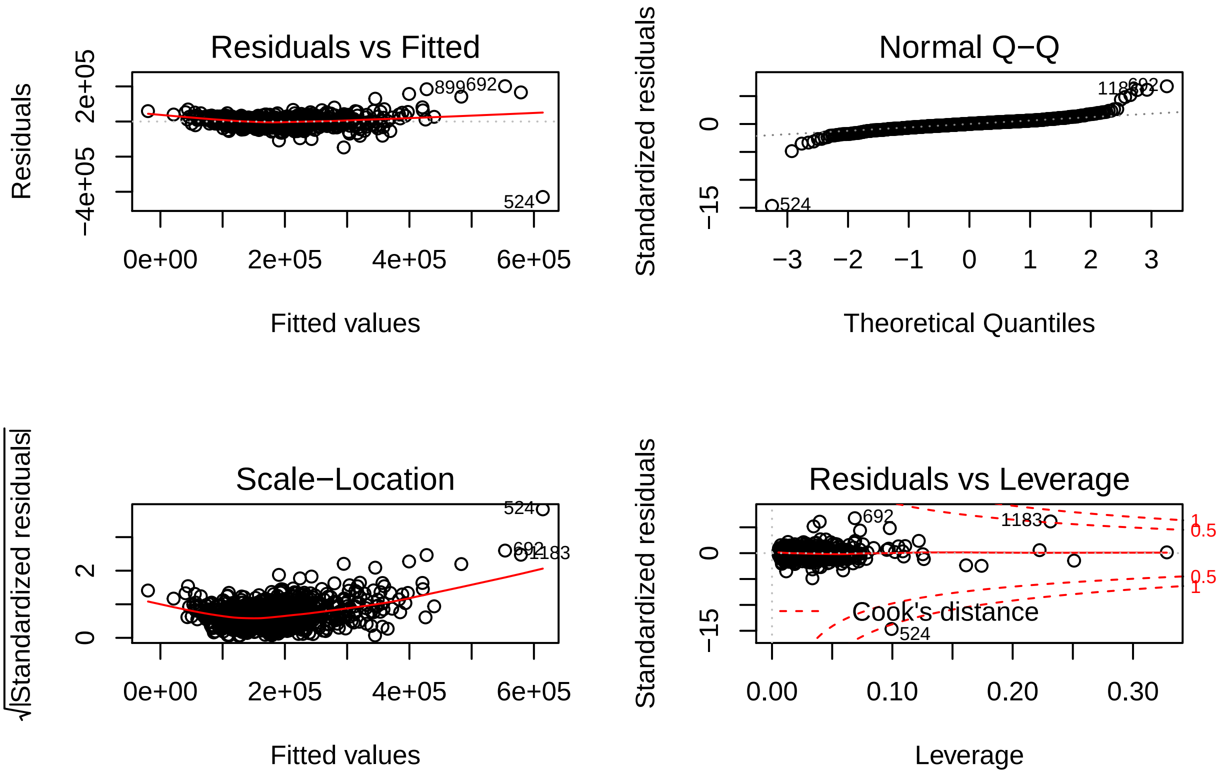
```r
ols_model2_rmse <- sqrt(mean(ols_model2$residuals^2))
ols_model2_rmse #30515.48, increases than the previous one
```

```
## [1] 30515.48
```

```r
model.apply <- function(model, testdata){
predict.test <- predict(model, testdata)
SSE <- sum((testdata$SalePrice - predict.test)^2)
SST <- sum((testdata$SalePrice -
mean(testdata$SalePrice))^2)
r.square <- 1-SSE/SST
test.rmse <- sqrt(mean((testdata$SalePrice - predict.test)^2))
par(mfrow = c(2,2))
plot(model)
return(c(r.square, test.rmse))
}
model.apply(ols_model2, testdata)
```

```
## Warning: contrasts dropped from factor BsmtQual
```

```
## Warning: contrasts dropped from factor KitchenQual
```




```
## [1] 7.528322e-01 3.933700e+04
```

```r
#rmse 39337.00, is high but the model fit is good
```

```r
library(car)
```

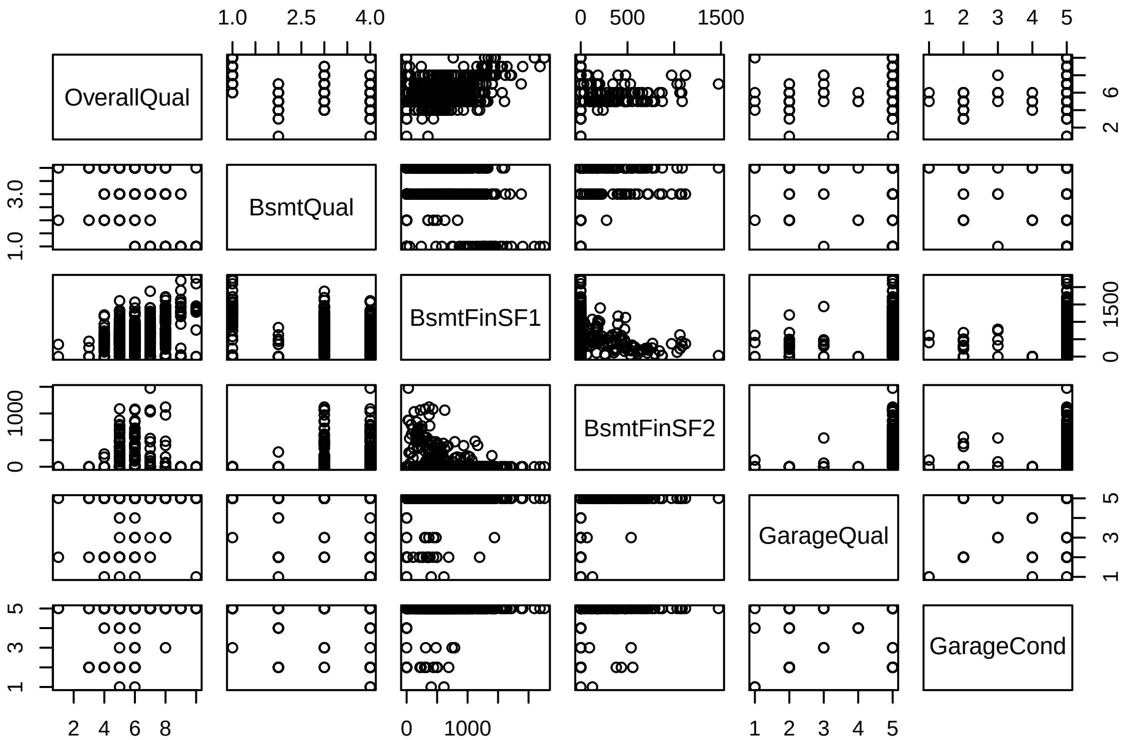
```
##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
vif(ols_model2) #check if there is one has vif >5
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## LotArea      1.349220  1        1.161559
## OverallQual  3.308102  1        1.818819
## OverallCond  1.357869  1        1.165276
## YearBuilt    3.312700  1        1.820082
## MasVnrArea   1.464371  1        1.210112
## BsmtQual     4.113350  3        1.265802
## BsmtFinSF1   4.311485  1        2.076412
## BsmtFinSF2   1.519670  1        1.232749
## BsmtUnfSF    4.133139  1        2.033012
## X1stFlrSF    4.023277  1        2.005811
## X2ndFlrSF    2.639135  1        1.624542
## KitchenQual  2.754232  3        1.183950
## KitchenAbvGr 1.206633  1        1.098469
## BedroomAbvGr 1.824756  1        1.350835
## GarageCars   2.007524  1        1.416871
```

```
## PoolArea      1.054093  1        1.026690
```

```r
# also check scatter plot
pairs(~ OverallQual+ BsmtQual +BsmtFinSF1 +
        BsmtFinSF2+GarageQual + GarageCond,
        data = traindata)
```




```r
ols_model3 <- lm(SalePrice ~ LotArea + OverallQual + OverallCond
                 + YearBuilt +  BsmtQual +BsmtFinSF1 +
                   BedroomAbvGr +X1stFlrSF +X2ndFlrSF
                 +KitchenQual + KitchenAbvGr + PoolArea
                 + OverallQual:GarageCars, data = traindata)
summary(ols_model3)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + OverallQual + OverallCond +
##     YearBuilt + BsmtQual + BsmtFinSF1 + BedroomAbvGr + X1stFlrSF +
##     X2ndFlrSF + KitchenQual + KitchenAbvGr + PoolArea + OverallQual:GarageCars,
##     data = traindata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -418097  -12189     407   12519  216474
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -6.242e+05  1.245e+05  -5.013 6.51e-07 ***
## LotArea                 8.629e-01  2.323e-01   3.715 0.000216 ***
## OverallQual             7.573e+03  1.581e+03   4.790 1.96e-06 ***
## OverallCond             6.674e+03  1.065e+03   6.266 5.84e-10 ***
## YearBuilt               3.404e+02  6.141e+01   5.544 3.95e-08 ***
```

```
## BsmtQual2             -3.940e+04  9.079e+03  -4.339 1.60e-05 ***
## BsmtQual3             -3.878e+04  5.055e+03  -7.671 4.62e-14 ***
## BsmtQual4             -4.212e+04  6.111e+03  -6.892 1.06e-11 ***
## BsmtFinSF1             2.153e+01  2.804e+00   7.679 4.37e-14 ***
## BedroomAbvGr          -3.543e+03  1.693e+03  -2.092 0.036701 *
## X1stFlrSF              7.035e+01  4.431e+00  15.878  < 2e-16 ***
## X2ndFlrSF              5.732e+01  3.652e+00  15.695  < 2e-16 ***
## KitchenQual2          -2.713e+04  8.485e+03  -3.197 0.001438 **
## KitchenQual3          -2.742e+04  5.200e+03  -5.273 1.70e-07 ***
## KitchenQual4          -3.259e+04  5.677e+03  -5.741 1.31e-08 ***
## KitchenAbvGr          -2.275e+04  5.189e+03  -4.385 1.30e-05 ***
## PoolArea               6.495e+01  2.343e+01   2.772 0.005689 **
## OverallQual:GarageCars  2.531e+03  3.404e+02   7.436 2.51e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30770 on 858 degrees of freedom
## Multiple R-squared:  0.8536, Adjusted R-squared:  0.8507
## F-statistic: 294.4 on 17 and 858 DF,  p-value: < 2.2e-16
```

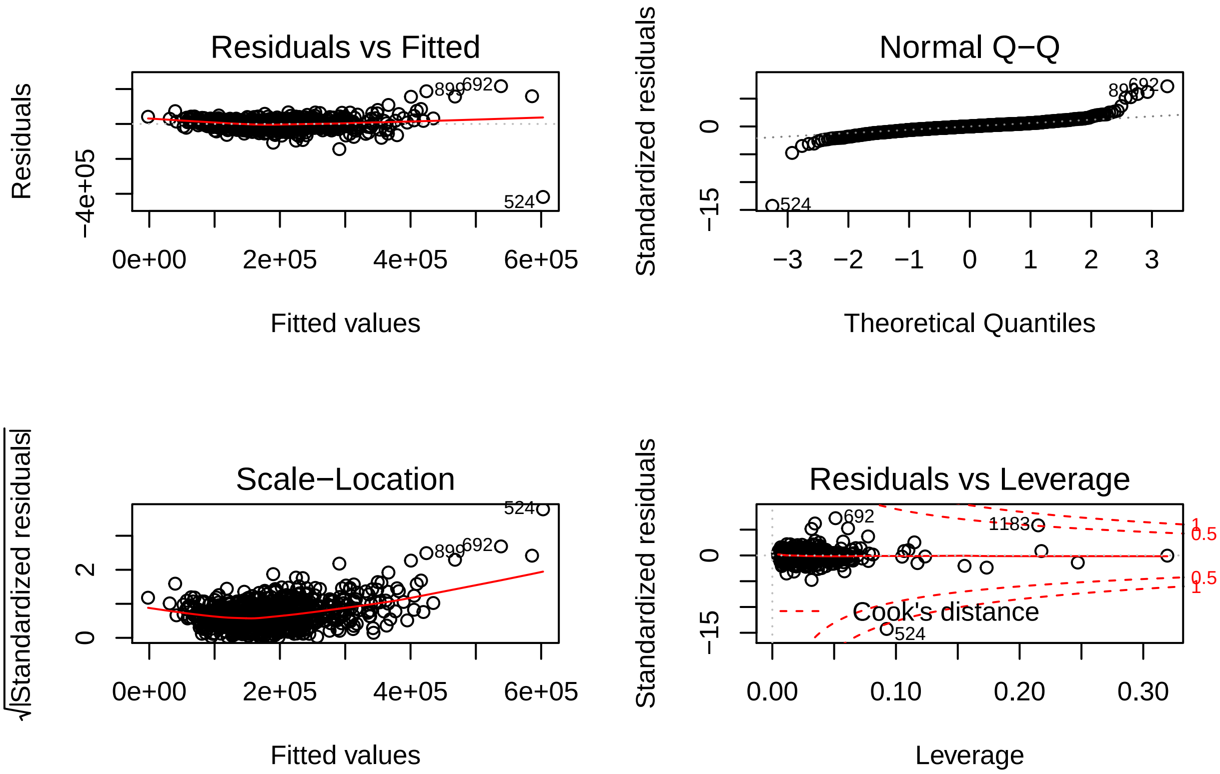
```r
vif(ols_model3)
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## LotArea                1.326730  1        1.151838
## OverallQual            4.138279  1        2.034276
## OverallCond            1.349792  1        1.161806
## YearBuilt              3.225588  1        1.795992
## BsmtQual               3.999655  3        1.259903
## BsmtFinSF1             1.362300  1        1.167176
## BedroomAbvGr           1.803438  1        1.342921
## X1stFlrSF              2.648837  1        1.627525
## X2ndFlrSF              2.432956  1        1.559794
## KitchenQual            2.675553  3        1.178245
## KitchenAbvGr           1.199346  1        1.095146
## PoolArea               1.040070  1        1.019838
## OverallQual:GarageCars 4.345762  1        2.084649
```

```r
model.apply(ols_model3, testdata)
```

```
## Warning: contrasts dropped from factor BsmtQual
```

```
## Warning: contrasts dropped from factor KitchenQual
```




```
## [1] 7.781851e-01 3.726496e+04
```

```r
#r^2: 77.82% rmse:37264.96; seems much better
```

```r
anova(ols_model2, ols_model3, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: SalePrice ~ LotArea + OverallQual + OverallCond + YearBuilt +
##     MasVnrArea + BsmtQual + BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF +
##     X1stFlrSF + X2ndFlrSF + KitchenQual + KitchenAbvGr + BedroomAbvGr +
##     GarageCars + PoolArea
## Model 2: SalePrice ~ LotArea + OverallQual + OverallCond + YearBuilt +
##     BsmtQual + BsmtFinSF1 + BedroomAbvGr + X1stFlrSF + X2ndFlrSF +
##     KitchenQual + KitchenAbvGr + PoolArea + OverallQual:GarageCars
##   Res.Df        RSS Df  Sum of Sq F Pr(>F)
## 1    855 8.1573e+11
## 2    858 8.1232e+11 -3 3404933732
```

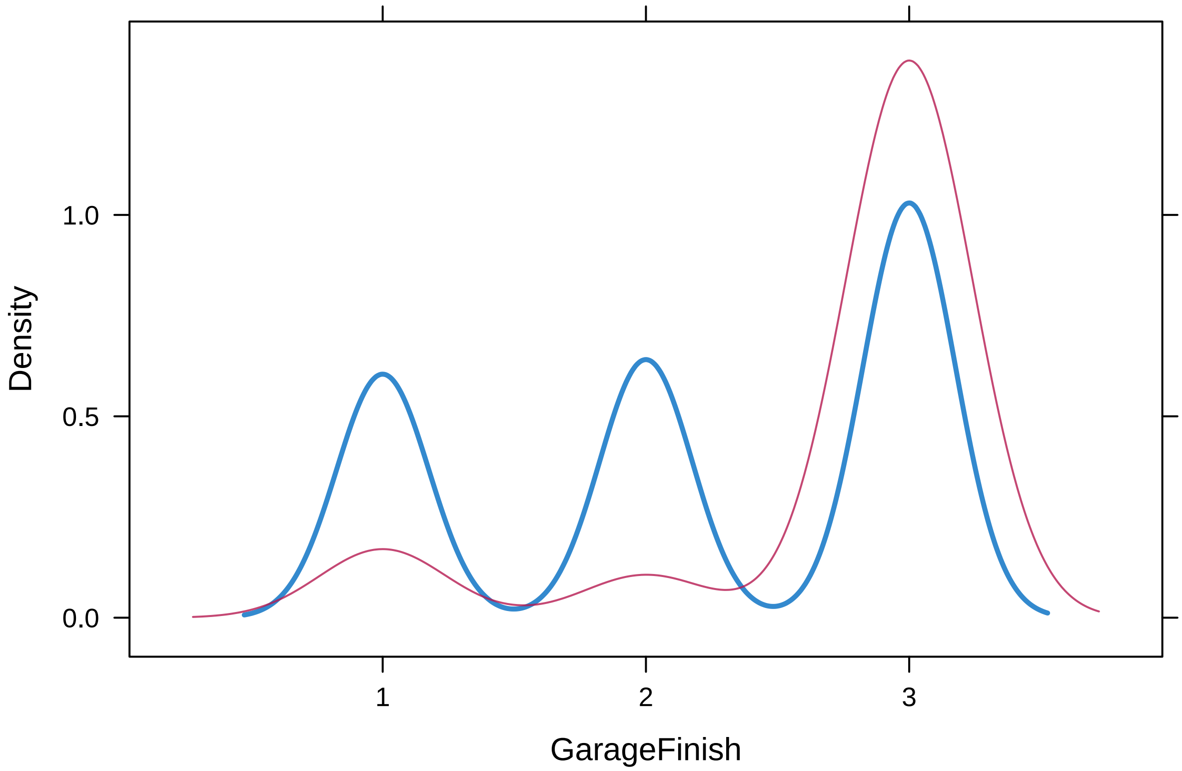
```r
library(dplyr)
data2 <- select(data2, -c(PoolQC, MiscFeature, Alley, Fence,
FireplaceQu, LotFrontage))
library(mice)
#using cart
imp_data2 <- mice(data2, m = 1, method = "cart", printFlag = FALSE)
table(imp_data2$imp$ GarageFinish)
```

```
##
## Fin RFn Unf
##   8   5  65
```

```
table(data2$ GarageFinish)
```

```
##
## Fin RFn Unf
## 367 389 625
```

```
densityplot(imp_data2, ~ GarageFinish) #from pattern it is acceptable
```




```
full_data2 <- complete(imp_data2)
# then double check no missing data
sort(sapply(full_data2, missing), decreasing = TRUE)[1:5] #no missing data
```

```
##   Utilities          Id  MSSubClass    MSZoning     LotArea
## 0.001370802 0.000000000 0.000000000 0.000000000 0.000000000
```

```
which(is.na(full_data2$Utilities))
```

```
## [1] 456 486
```

```
full_data2$Utilities[c(456,486)] <- 'AllPub'
missing(full_data2$Utilities)
```

```
## [1] 0
```

```
full_data2no <- full_data2[, -1]
data1_x <- full_data1[, -74]
dim(data1_x)
```

```
## [1] 1460   73
```

```
dim(full_data2no)
```

```
## [1] 1459   73
```

```r
comb <- rbind(data1_x, full_data2no)
mat <- model.matrix(~., data = comb)[,-1]
data1.matrix <- mat[1:1460,]
data2.matrix <- mat[1461:2919, ]
train.mat <- data1.matrix[train, ]
test.mat <- data1.matrix[test, ]
dim(mat)
```

```
## [1] 2919  231
```

```r
y <- traindata$SalePrice
#use package
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 3.4.2
```

```
## Loading required package: Matrix
```

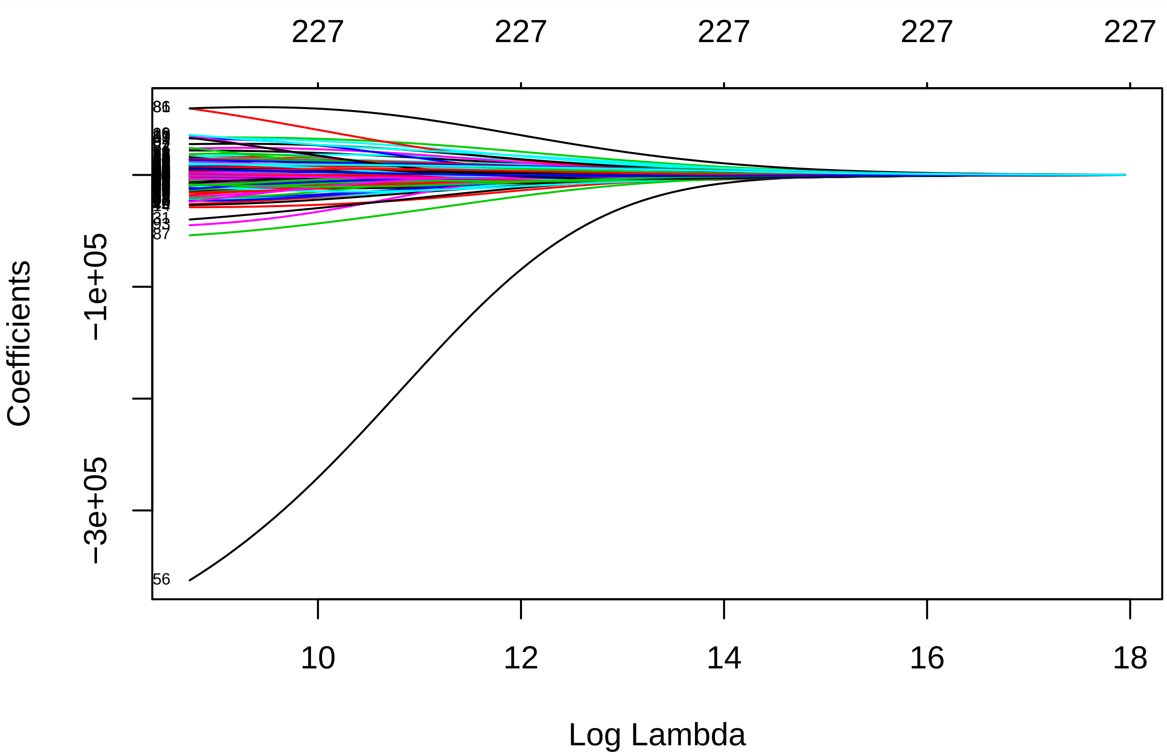
```
## Warning: package 'Matrix' was built under R version 3.4.2
```

```
## Loading required package: foreach
```

```
## Warning: package 'foreach' was built under R version 3.4.3
```

```
## Loaded glmnet 2.0-13
```

```r
ridge_model <- glmnet(train.mat, y, alpha = 0)
plot(ridge_model, xvar = "lambda", label = TRUE)
```




```r
# this is just for [visualize]
#this will give us the optimal lambda
set.seed(11)
ridge_model2 <- cv.glmnet(train.mat, y, alpha = 0)
# cv.glmnet uses cross-validate to find lambda
```

```
lambda <-ridge_model2$lambda.min
lambda #20870.61
```

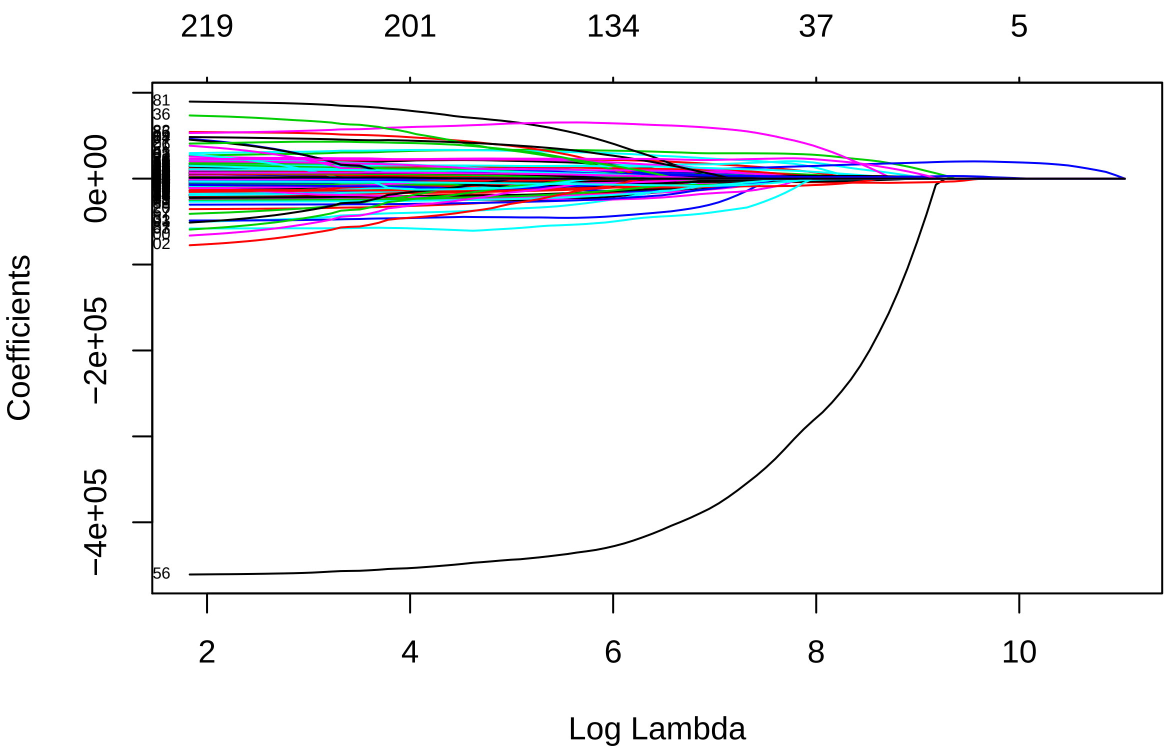
```
## [1] 20870.61
```

```
#what are the coeff?
ridge_coe <-predict(ridge_model, train.mat, s = lambda,
type = "coefficient")
#then apply the model to test data
y.test <- testdata$SalePrice
ridge_y.test.predict <- predict(ridge_model, test.mat,
                                s = lambda)
ridge_model_rmse <- sqrt(mean((y.test - ridge_y.test.predict)^2))
ridge_model_rmse #37093.79
```

```
## [1] 37093.79
```

Apply Lasso

```
#same things here, first visualize
lasso_model <- glmnet(train.mat, y, alpha = 1)
plot(lasso_model, xvar = "lambda", label = TRUE)
```




```
set.seed(11)
lasso_model2 <- cv.glmnet(train.mat, y, alpha = 1)
lambda_lasso <- lasso_model2$lambda.min
lasso_y.test.predict <- predict(lasso_model, newx = test.mat,
                                s = lambda_lasso)
lasso_model_rmse <- sqrt(mean((lasso_y.test.predict - y.test)^2))
lasso_model_rmse #39717.59
```

```
## [1] 39717.59
```

Ridge gives the smallest rmse, use ridge to predict

```
BestP<- data.frame(Id = data2$Id,
                       SalePrice =predict(ridge_model,
                                          data2.matrix,
                                           s = lambda))
write.csv(BestP, "BestP.csv", row.names = FALSE)
```