Data Analysis - Ames House Prices

Random85

2023-07-20

The Dataset

```r
library(faraway)
library(dplyr)
library(psych)
library(corrplot)
library(ggplot2)
library(ggcorrplot)
library(lares)
library(reshape2)

housing_data = read.csv("dataset\\AmesHousing.csv")

# summary(housing_data)

nrow(housing_data)
```

```
## [1] 2930
```

```r
ncol(housing_data)
```

```
## [1] 82
```

```r
# Using small chunk of dataset for quick testing :)

#housing_data = housing_data[1:100,]

# Coercing categorical predictors into factor variables

housing_data[is.na(housing_data)] = 1

for (i in 1:ncol(housing_data)) {
  if (typeof(housing_data[, i]) == "character") {
    if (length(unique(housing_data[, i])) >= 2) {
      housing_data[, i] = as.factor(housing_data[, i])
    }

  }
```

```
}

housing_data$Yr.Sold = as.factor(housing_data$Yr.Sold)

str(housing_data)
```

```
## 'data.frame':    2930 obs. of  82 variables:
##  $ Order          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ PID            : int  526301100 526350040 526351010 526353030 527105010 527105030 527127150 527145
##  $ MS.SubClass    : int  20 20 20 20 60 60 120 120 120 60 ...
##  $ MS.Zoning      : Factor w/ 7 levels "A (agr)","C (all)",..: 6 5 6 6 6 6 6 6 6 6 ...
##  $ Lot.Frontage   : num  141 80 81 93 74 78 41 43 39 60 ...
##  $ Lot.Area       : int  31770 11622 14267 11160 13830 9978 4920 5005 5389 7500 ...
##  $ Street         : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Alley          : Factor w/ 3 levels "1","Grvl","Pave": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Lot.Shape      : Factor w/ 4 levels "IR1","IR2","IR3",..: 1 4 1 4 1 1 4 1 1 4 ...
##  $ Land.Contour   : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 2 4 4 ...
##  $ Utilities      : Factor w/ 3 levels "AllPub","NoSeWa",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Lot.Config     : Factor w/ 5 levels "Corner","CulDSac",..: 1 5 1 1 5 5 5 5 5 5 ...
##  $ Land.Slope     : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Neighborhood   : Factor w/ 28 levels "Blmngtn","Blueste",..: 16 16 16 16 9 9 25 25 25 9 ...
##  $ Condition.1    : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 3 3 3 ...
##  $ Condition.2    : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Bldg.Type      : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 5 5 5 1 ...
##  $ House.Style    : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 3 3 3 3 6 6 3 3 3 6 ...
##  $ Overall.Qual   : int  6 5 6 7 5 6 8 8 8 7 ...
##  $ Overall.Cond   : int  5 6 6 5 5 6 5 5 5 5 ...
##  $ Year.Built     : int  1960 1961 1958 1968 1997 1998 2001 1992 1995 1999 ...
##  $ Year.Remod.Add : int  1960 1961 1958 1968 1998 1998 2001 1992 1996 1999 ...
##  $ Roof.Style     : Factor w/ 6 levels "Flat","Gable",..: 4 2 4 4 2 2 2 2 2 2 ...
##  $ Roof.Matl      : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Exterior.1st   : Factor w/ 16 levels "AsbShng","AsphShn",..: 4 14 15 4 14 14 6 7 6 14 ...
##  $ Exterior.2nd   : Factor w/ 17 levels "AsbShng","AsphShn",..: 11 15 16 4 15 15 6 7 6 15 ...
##  $ Mas.Vnr.Type   : Factor w/ 6 levels "","BrkCmn","BrkFace",..: 6 5 3 5 5 3 5 5 5 5 ...
##  $ Mas.Vnr.Area   : num  112 0 108 0 0 20 0 0 0 0 ...
##  $ Exter.Qual     : Factor w/ 4 levels "Ex","Fa","Gd",..: 4 4 4 3 4 4 3 3 3 4 ...
##  $ Exter.Cond     : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Foundation     : Factor w/ 6 levels "BrkTil","CBlock",..: 2 2 2 2 3 3 3 3 3 3 ...
##  $ Bsmt.Qual      : Factor w/ 7 levels "","1","Ex","Fa",..: 7 7 7 7 5 7 5 5 5 7 ...
##  $ Bsmt.Cond      : Factor w/ 7 levels "","1","Ex","Fa",..: 5 7 7 7 7 7 7 7 7 7 ...
##  $ Bsmt.Exposure  : Factor w/ 6 levels "","1","Av","Gd",..: 4 6 6 6 6 6 5 6 6 6 ...
##  $ BsmtFin.Type.1 : Factor w/ 8 levels "","1","ALQ","BLQ",..: 4 7 3 3 5 5 5 3 5 8 ...
##  $ BsmtFin.SF.1   : num  639 468 923 1065 791 ...
##  $ BsmtFin.Type.2 : Factor w/ 8 levels "","1","ALQ","BLQ",..: 8 6 8 8 8 8 8 8 8 8 ...
##  $ BsmtFin.SF.2   : num  0 144 0 0 0 0 0 0 0 0 ...
##  $ Bsmt.Unf.SF    : num  441 270 406 1045 137 ...
##  $ Total.Bsmt.SF  : num  1080 882 1329 2110 928 ...
##  $ Heating        : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Heating.QC     : Factor w/ 5 levels "Ex","Fa","Gd",..: 2 5 5 1 3 1 1 1 1 3 ...
##  $ Central.Air    : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Electrical     : Factor w/ 6 levels "","FuseA","FuseF",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ X1st.Flr.SF    : int  1656 896 1329 2110 928 926 1338 1280 1616 1028 ...
##  $ X2nd.Flr.SF    : int  0 0 0 0 701 678 0 0 0 776 ...
```

```
##  $ Low.Qual.Fin.SF: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Gr.Liv.Area    : int  1656 896 1329 2110 1629 1604 1338 1280 1616 1804 ...
##  $ Bsmt.Full.Bath : num  1 0 0 1 0 0 1 0 1 0 ...
##  $ Bsmt.Half.Bath : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Full.Bath      : int  1 1 1 2 2 2 2 2 2 2 ...
##  $ Half.Bath      : int  0 0 1 1 1 1 0 0 0 1 ...
##  $ Bedroom.AbvGr  : int  3 2 3 3 3 3 2 2 2 3 ...
##  $ Kitchen.AbvGr  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Kitchen.Qual   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 3 1 5 3 3 3 3 3 ...
##  $ TotRms.AbvGrd  : int  7 5 6 8 6 7 6 5 5 7 ...
##  $ Functional     : Factor w/ 8 levels "Maj1","Maj2",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ Fireplaces     : int  2 0 0 2 1 1 0 0 1 1 ...
##  $ Fireplace.Qu   : Factor w/ 6 levels "1","Ex","Fa",..: 4 1 1 6 6 4 1 1 6 6 ...
##  $ Garage.Type    : Factor w/ 7 levels "1","2Types","Attchd",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Garage.Yr.Blt  : num  1960 1961 1958 1968 1997 ...
##  $ Garage.Finish  : Factor w/ 5 levels "","1","Fin","RFn",..: 3 5 5 3 3 3 3 4 4 3 ...
##  $ Garage.Cars    : num  2 1 1 2 2 2 2 2 2 2 ...
##  $ Garage.Area    : num  528 730 312 522 482 470 582 506 608 442 ...
##  $ Garage.Qual    : Factor w/ 7 levels "","1","Ex","Fa",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ Garage.Cond    : Factor w/ 7 levels "","1","Ex","Fa",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ Paved.Drive    : Factor w/ 3 levels "N","P","Y": 2 3 3 3 3 3 3 3 3 3 ...
##  $ Wood.Deck.SF   : int  210 140 393 0 212 360 0 0 237 140 ...
##  $ Open.Porch.SF  : int  62 0 36 0 34 36 0 82 152 60 ...
##  $ Enclosed.Porch : int  0 0 0 0 0 0 170 0 0 0 ...
##  $ X3Ssn.Porch    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Screen.Porch   : int  0 120 0 0 0 0 0 144 0 0 ...
##  $ Pool.Area      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Pool.QC        : Factor w/ 5 levels "1","Ex","Fa",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Fence          : Factor w/ 5 levels "1","GdPrv","GdWo",..: 1 4 1 1 4 1 1 1 1 1 ...
##  $ Misc.Feature   : Factor w/ 6 levels "1","Elev","Gar2",..: 1 1 3 1 1 1 1 1 1 1 ...
##  $ Misc.Val       : int  0 0 12500 0 0 0 0 0 0 0 ...
##  $ Mo.Sold        : int  5 6 6 4 3 6 4 1 3 6 ...
##  $ Yr.Sold        : Factor w/ 5 levels "2006","2007",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Sale.Type      : Factor w/ 10 levels "COD","Con","ConLD",..: 10 10 10 10 10 10 10 10 10 10 ...
##  $ Sale.Condition : Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ SalePrice      : int  215000 105000 172000 244000 189900 195500 213500 191500 236500 189000 ...
```

First few examples:

```r
housing_data$SalePrice[1:10]
```

```
##  [1] 215000 105000 172000 244000 189900 195500 213500 191500 236500 189000
```

```r
housing_data$Lot.Area[1:10]
```

```
##  [1] 31770 11622 14267 11160 13830  9978  4920  5005  5389  7500
```

```r
housing_data$Utilities[1:10]
```

```
##  [1] AllPub AllPub AllPub AllPub AllPub AllPub AllPub AllPub AllPub AllPub
## Levels: AllPub NoSeWa NoSewr
```

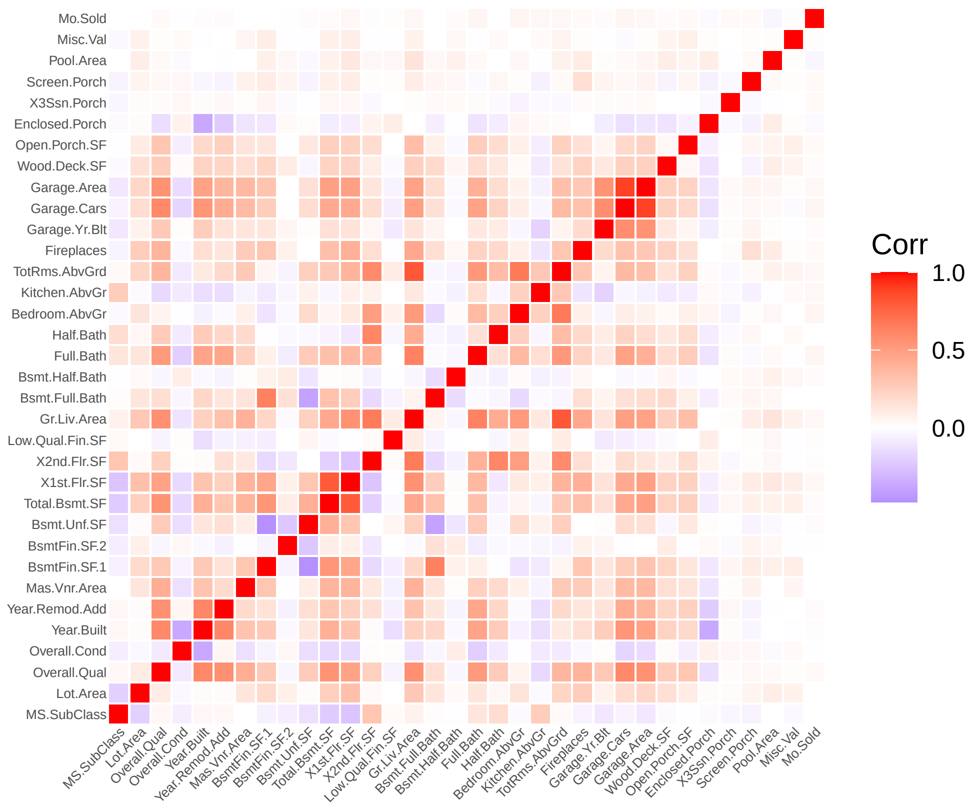
Dropping some of the columns that are not useful as a predictor for sale price.

```
# Dropping the order and PID, looks like this is just for record keeping
housing_data = subset(housing_data, select = -c(Order, PID))

# Removing lot fraontage area, since we have lot area. Frontage is measurement of the house start to th
housing_data = subset(housing_data, select = -c(Lot.Frontage))

# Removing alley because we have street data
housing_data = subset(housing_data, select = -c(Alley))

# removing one condition column and exterior
housing_data = subset(housing_data, select = -c(Condition.2, Exterior.2nd))
```

```
View(housing_data)
```

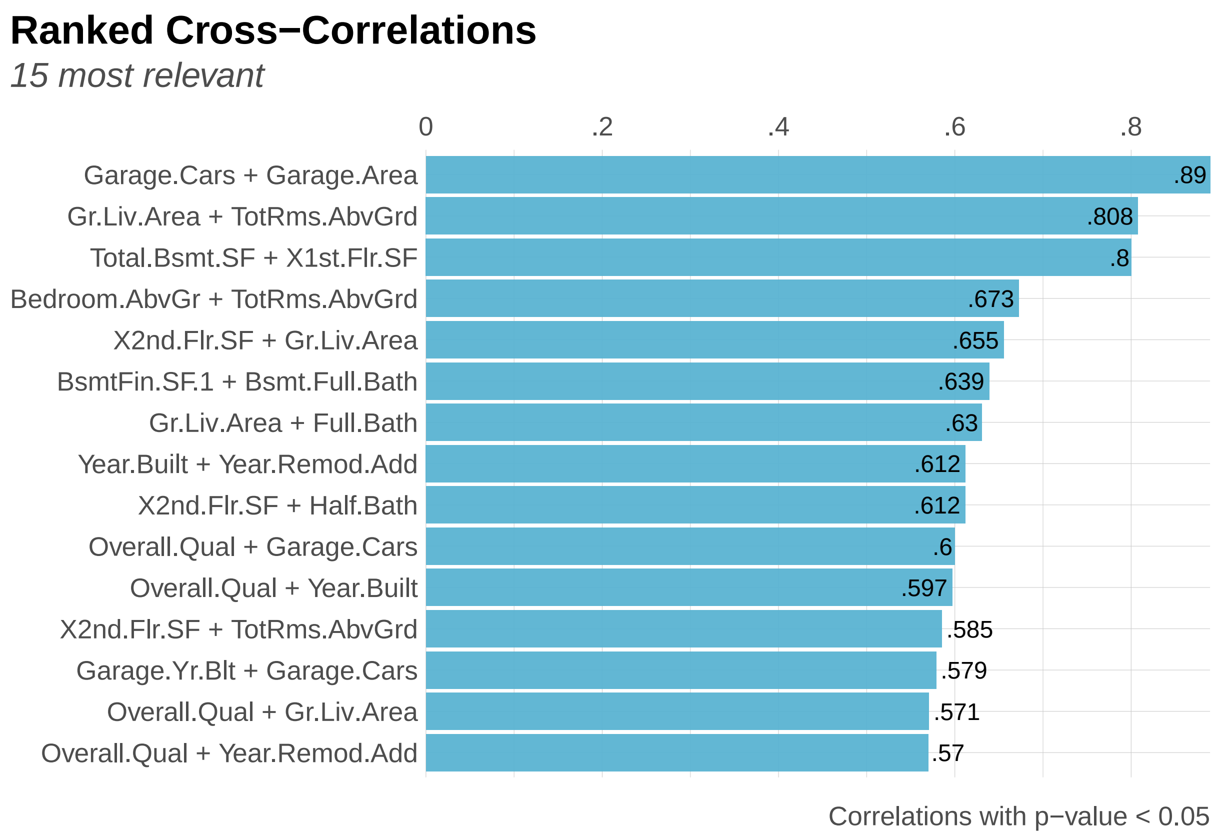
Collinearity and correlation analysis

```
# Subsetting all the numeric elements of the dataset for collinearity and correlation analysis:

n_idxs = unlist(lapply(housing_data, is.numeric), use.names = FALSE)
all_numeric_housing_data = housing_data[, n_idxs]
nhd = all_numeric_housing_data = housing_data[, n_idxs]
numeric_housing_data = subset(all_numeric_housing_data, select = -c(SalePrice))

#str(numeric_housing_data)



# MUST specify use = "complete.obs" argument to ignore NA's in dataset
corrs = round(cor(numeric_housing_data, use="complete.obs"), 2)


# some possible correlation plots?

#corrplot(corrs, method="number")
#ggcorrplot(corrs, lab_size = 0.1)

#corrs

ggplot(melt(corrs), aes(Var1, Var2, fill=value)) +
  geom_tile(height=0.9, width=0.9) +
  scale_fill_gradient2(low="blue", mid="white", high="red") +
  theme_minimal() +
  coord_equal() +
  labs(x="",y="",fill="Corr") +
  theme(axis.text.x=element_text(size=5, angle=45, vjust=1, hjust=1,
                                 margin=margin(-3,0,0,0)),
        axis.text.y=element_text(size=5, margin=margin(0,-3,0,0)),
        panel.grid.major=element_blank())
```



Taking a closer look at some of the most correlated predictors:

```
corr_cross(numeric_housing_data,
  max_pvalue = 0.05,
  top = 15
)
```




```
# A few of these further visualized in pairs:
names(numeric_housing_data)
```

```
##  [1] "MS.SubClass"     "Lot.Area"        "Overall.Qual"    "Overall.Cond"
##  [5] "Year.Built"      "Year.Remod.Add"  "Mas.Vnr.Area"    "BsmtFin.SF.1"
##  [9] "BsmtFin.SF.2"    "Bsmt.Unf.SF"     "Total.Bsmt.SF"   "X1st.Flr.SF"
## [13] "X2nd.Flr.SF"     "Low.Qual.Fin.SF" "Gr.Liv.Area"     "Bsmt.Full.Bath"
## [17] "Bsmt.Half.Bath"  "Full.Bath"       "Half.Bath"       "Bedroom.AbvGr"
## [21] "Kitchen.AbvGr"   "TotRms.AbvGrd"   "Fireplaces"      "Garage.Yr.Blt"
## [25] "Garage.Cars"     "Garage.Area"     "Wood.Deck.SF"    "Open.Porch.SF"
## [29] "Enclosed.Porch"  "X3Ssn.Porch"     "Screen.Porch"    "Pool.Area"
## [33] "Misc.Val"        "Mo.Sold"
```

```
# pairs(numeric_housing_data, col = "dodgerblue")
```

While some of these high correlation measures are to be expected, such as house year built along garage year built, we can also see some non-trivial patterns start to emerge from the more continuous numeric predictors.

Models (VP)

```
#Functions to evaluate models
library(lmtest)
```

```
get_bp_decision = function(model, alpha) {
  decide = unname(bptest(model)$p.value < alpha)
  ifelse(decide, "Reject", "Fail to Reject")
}

get_sw_decision = function(model, alpha) {
  decide = unname(shapiro.test(resid(model))$p.value < alpha)
  ifelse(decide, "Reject", "Fail to Reject")
}

get_num_params = function(model) {
  length(coef(model))
}

get_loocv_rmse = function(model) {
  sqrt(mean((resid(model) / (1 - hatvalues(model))) ^ 2,na.rm=TRUE))
}

get_adj_r2 = function(model) {
  summary(model)$adj.r.squared
}
```

Location of the house could play a big role in house price, but for the dataset that is used, it has total 28 neighbors as follow:

```
length(unique(housing_data[, "Neighborhood"]))
```

```
## [1] 28
```

We don't need this big list of dummy variable, so creating new variable as location based on some important factor

```
# first converting the some exterior variables to numeric and then using it:
levels(housing_data[,"Exter.Cond"])
```

```
## [1] "Ex" "Fa" "Gd" "Po" "TA"
```

```
levels(housing_data[,"Exter.Qual"])
```

```
## [1] "Ex" "Fa" "Gd" "TA"
```

```
levels(housing_data[,"Functional"])
```

```
## [1] "Maj1" "Maj2" "Min1" "Min2" "Mod"  "Sal"  "Sev"  "Typ"
```

```r
housing_data$Exter.Cond.Num = 5 - as.numeric(housing_data[, "Exter.Cond"])
housing_data$Exter.Qual.Num = 4 - as.numeric(housing_data[, "Exter.Qual"])
housing_data$Functional.Num = ifelse(housing_data[,"Functional"] == "Maj1", 8,
                                     ifelse(housing_data[,"Functional"] == "Maj2", 7,
                                            ifelse(housing_data[,"Functional"] == "Min1", 6,
                                                   ifelse(housing_data[,"Functional"] == "Min2", 5,
                                                          ifelse(housing_data[,"Functional"] == "Mod", 4
                                                                 ifelse(housing_data[,"Functional"] ==
                                                                        ifelse(housing_data[,"Functiona
                                                                               ifelse(housing_data[,"Fu

housing_data$Location = (housing_data[,"Overall.Qual"] / mean(housing_data[,"Overall.Qual"])) +
                        (housing_data[,"Overall.Cond"] / mean(housing_data[,"Overall.Cond"])) +
                        (housing_data[,"Exter.Cond.Num"] / mean(housing_data[,"Exter.Cond.Num"])) +
                        (housing_data[,"Exter.Qual.Num"] / mean(housing_data[,"Exter.Qual.Num"])) +
                        (housing_data[,"Functional.Num"] / mean(housing_data[,"Functional.Num"]))

summary(housing_data[,"Location"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.631   2.846   3.613   5.000   5.281  24.363
```

Make sure that location is corelated to saleprice and positive:

```r
cor(housing_data[,c("Location","SalePrice")])
```

```
##            Location SalePrice
## Location  1.0000000 0.2533568
## SalePrice 0.2533568 1.0000000
```

It has weak but a positive relation of with saleprice.

Taking most relevant predictors that are logically be useful to build saleprice model with existing knowledge of real estate.

```r
# this model removes some of the qualities variable because it has some condition of those variables pr
# Such as, keeping Bsmt.Cond and removing Bsmt.Qual (because both factor has almost same levels)

saleprice_full_model_selected = lm(SalePrice ~ MS.Zoning + Lot.Area + Street + Lot.Shape + Land.Contour

saleprice_additive_model = lm(SalePrice ~ ., data = housing_data)

saleprice_selected_backward_aic = step(saleprice_full_model_selected, direction = "backward", trace = 0
```

```r
# check the number of predictors in the model
length(coef(saleprice_full_model_selected)) - 1
```

```
## [1] 112
```

```r
length(coef(saleprice_selected_backward_aic)) - 1
```

```
## [1] 85
```

Testing the built model

```r
get_loocv_rmse(saleprice_additive_model)
```

```
## [1] Inf
```

```r
get_loocv_rmse(saleprice_full_model_selected)
```

```
## [1] Inf
```

```r
get_loocv_rmse(saleprice_selected_backward_aic)
```

```
## [1] Inf
```

Even though removing so many variables, the model is still very huge, so, removing some of the variables that are may be corelated. Also, removing the variables that seems related based on the common real estate knowledge.

```r
saleprice_full_model_selected_reduced = lm(SalePrice ~ MS.Zoning + Lot.Area + Street + Lot.Shape + Util

saleprice_selected_reduced_backward_aic = step(saleprice_full_model_selected_reduced, direction = "backw

summary(saleprice_selected_reduced_backward_aic)
```

```
##
## Call:
## lm(formula = SalePrice ~ MS.Zoning + Lot.Area + Street + Lot.Shape +
##     Land.Slope + House.Style + Year.Built + Foundation + Location +
##     Central.Air + Gr.Liv.Area + Bedroom.AbvGr + Kitchen.AbvGr +
##     Kitchen.Qual + TotRms.AbvGrd + Fireplaces + Garage.Area +
##     Wood.Deck.SF + Pool.Area + Misc.Val, data = housing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -468197  -16897    -685   14614  258882
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -9.777e+05  8.353e+04 -11.704  < 2e-16 ***
## MS.ZoningC (all)   7.126e+04  2.567e+04   2.776 0.005539 **
## MS.ZoningFV        8.987e+04  2.484e+04   3.618 0.000302 ***
## MS.ZoningI (all)   3.702e+04  3.522e+04   1.051 0.293332
## MS.ZoningRH        8.071e+04  2.550e+04   3.165 0.001566 **
```

```
## MS.ZoningRL        8.792e+04  2.463e+04   3.570 0.000363 ***
## MS.ZoningRM        8.265e+04  2.470e+04   3.346 0.000831 ***
## Lot.Area           5.945e-01  1.083e-01   5.488 4.43e-08 ***
## StreetPave         1.770e+04  1.082e+04   1.636 0.101926
## Lot.ShapeIR2       3.321e+03  4.203e+03   0.790 0.429420
## Lot.ShapeIR3      -4.011e+04  8.936e+03  -4.489 7.44e-06 ***
## Lot.ShapeReg      -4.391e+03  1.472e+03  -2.983 0.002879 **
## Land.SlopeMod      1.313e+04  3.222e+03   4.074 4.74e-05 ***
## Land.SlopeSev     -1.904e+04  1.004e+04  -1.897 0.057947 .
## House.Style1.5Unf  2.538e+04  8.226e+03   3.085 0.002055 **
## House.Style1Story  1.503e+04  2.546e+03   5.905 3.94e-09 ***
## House.Style2.5Fin -3.074e+04  1.266e+04  -2.428 0.015222 *
## House.Style2.5Unf  1.337e+04  7.525e+03   1.776 0.075821 .
## House.Style2Story -6.569e+03  2.540e+03  -2.586 0.009750 **
## House.StyleSFoyer  1.498e+04  4.634e+03   3.232 0.001242 **
## House.StyleSLvl    1.273e+03  3.893e+03   0.327 0.743669
## Year.Built         5.024e+02  4.058e+01  12.380  < 2e-16 ***
## FoundationCBlock  -5.768e+03  2.687e+03  -2.147 0.031907 *
## FoundationPConc    6.146e+03  3.191e+03   1.926 0.054201 .
## FoundationSlab    -2.737e+04  5.686e+03  -4.814 1.55e-06 ***
## FoundationStone    2.347e+04  1.080e+04   2.174 0.029787 *
## FoundationWood    -1.379e+04  1.578e+04  -0.874 0.382393
## Location           1.304e+03  2.233e+02   5.841 5.77e-09 ***
## Central.AirY       5.283e+03  3.008e+03   1.757 0.079091 .
## Gr.Liv.Area        7.567e+01  2.829e+00  26.750  < 2e-16 ***
## Bedroom.AbvGr     -6.832e+03  1.162e+03  -5.877 4.64e-09 ***
## Kitchen.AbvGr     -2.768e+04  3.499e+03  -7.911 3.60e-15 ***
## Kitchen.QualFa    -7.603e+04  5.402e+03 -14.073  < 2e-16 ***
## Kitchen.QualGd    -6.464e+04  2.837e+03 -22.784  < 2e-16 ***
## Kitchen.QualPo    -5.936e+04  3.471e+04  -1.710 0.087322 .
## Kitchen.QualTA    -7.565e+04  3.270e+03 -23.136  < 2e-16 ***
## TotRms.AbvGrd      2.254e+03  8.487e+02   2.656 0.007953 **
## Fireplaces         9.703e+03  1.194e+03   8.130 6.29e-16 ***
## Garage.Area        4.512e+01  3.955e+00  11.406  < 2e-16 ***
## Wood.Deck.SF       2.304e+01  5.458e+00   4.221 2.51e-05 ***
## Pool.Area         -5.825e+01  1.831e+01  -3.182 0.001480 **
## Misc.Val          -8.410e+00  1.136e+00  -7.406 1.70e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34320 on 2888 degrees of freedom
## Multiple R-squared:  0.8181, Adjusted R-squared:  0.8155
## F-statistic: 316.7 on 41 and 2888 DF,  p-value: < 2.2e-16
```

```r
# temp
length(coef(saleprice_full_model_selected_reduced)) - 1
```

```
## [1] 65
```

```r
length(coef(saleprice_selected_reduced_backward_aic)) - 1
```

```
## [1] 41
```

Testing the built model

```
get_loocv_rmse(saleprice_additive_model)
```

```
## [1] Inf
```

```
get_loocv_rmse(saleprice_full_model_selected_reduced)
```

```
## [1] Inf
```

```
get_loocv_rmse(saleprice_selected_reduced_backward_aic)
```

```
## [1] 35945.44
```

```
anova(saleprice_full_model_selected_reduced, saleprice_selected_reduced_backward_aic)
```

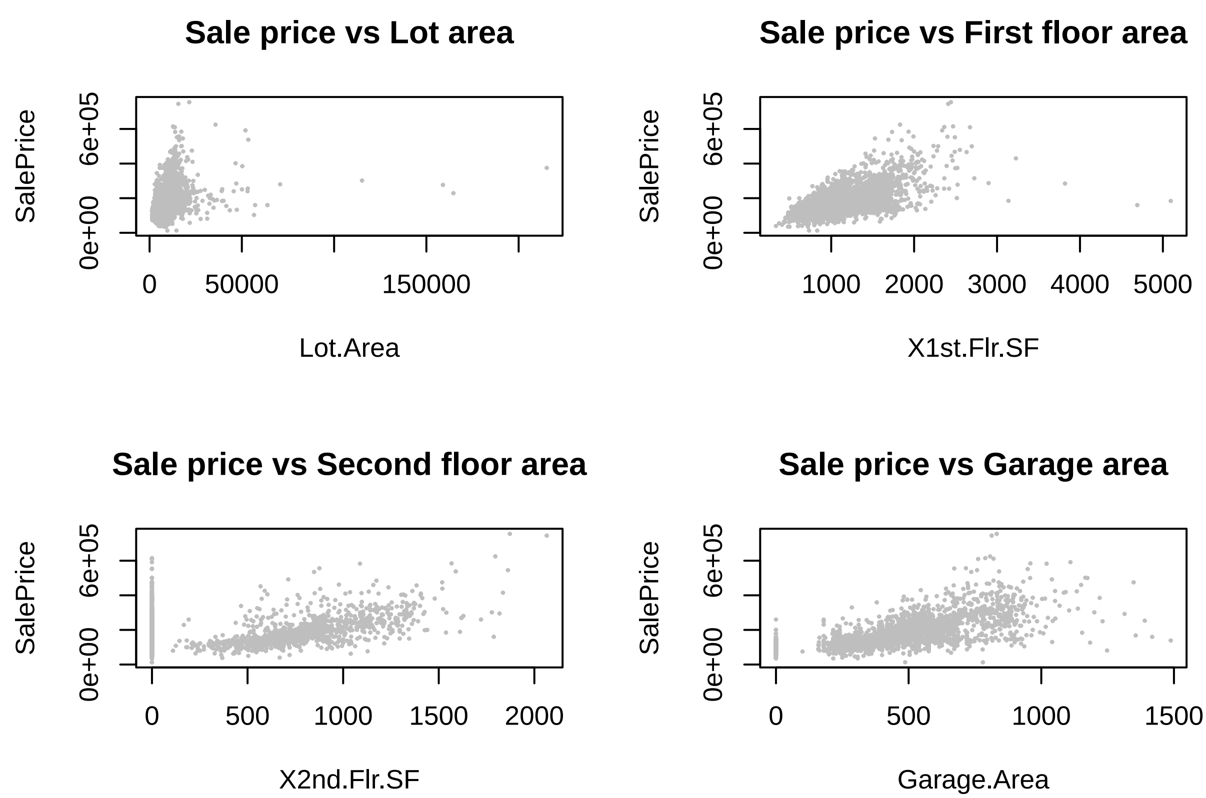
```
## Analysis of Variance Table
##
## Model 1: SalePrice ~ MS.Zoning + Lot.Area + Street + Lot.Shape + Utilities +
##     Land.Slope + House.Style + Year.Built + Foundation + Location +
##     Heating + Central.Air + Electrical + Gr.Liv.Area + Full.Bath +
##     Bedroom.AbvGr + Kitchen.AbvGr + Kitchen.Qual + TotRms.AbvGrd +
##     Fireplaces + Garage.Area + Paved.Drive + Wood.Deck.SF + Open.Porch.SF +
##     Pool.Area + Fence + Misc.Val + Yr.Sold
## Model 2: SalePrice ~ MS.Zoning + Lot.Area + Street + Lot.Shape + Land.Slope +
##     House.Style + Year.Built + Foundation + Location + Central.Air +
##     Gr.Liv.Area + Bedroom.AbvGr + Kitchen.AbvGr + Kitchen.Qual +
##     TotRms.AbvGrd + Fireplaces + Garage.Area + Wood.Deck.SF +
##     Pool.Area + Misc.Val
##   Res.Df        RSS  Df   Sum of Sq      F Pr(>F)
## 1   2864 3.3783e+12
## 2   2888 3.4010e+12 -24 -2.2687e+10 0.8014 0.7389
```

Variable transformations

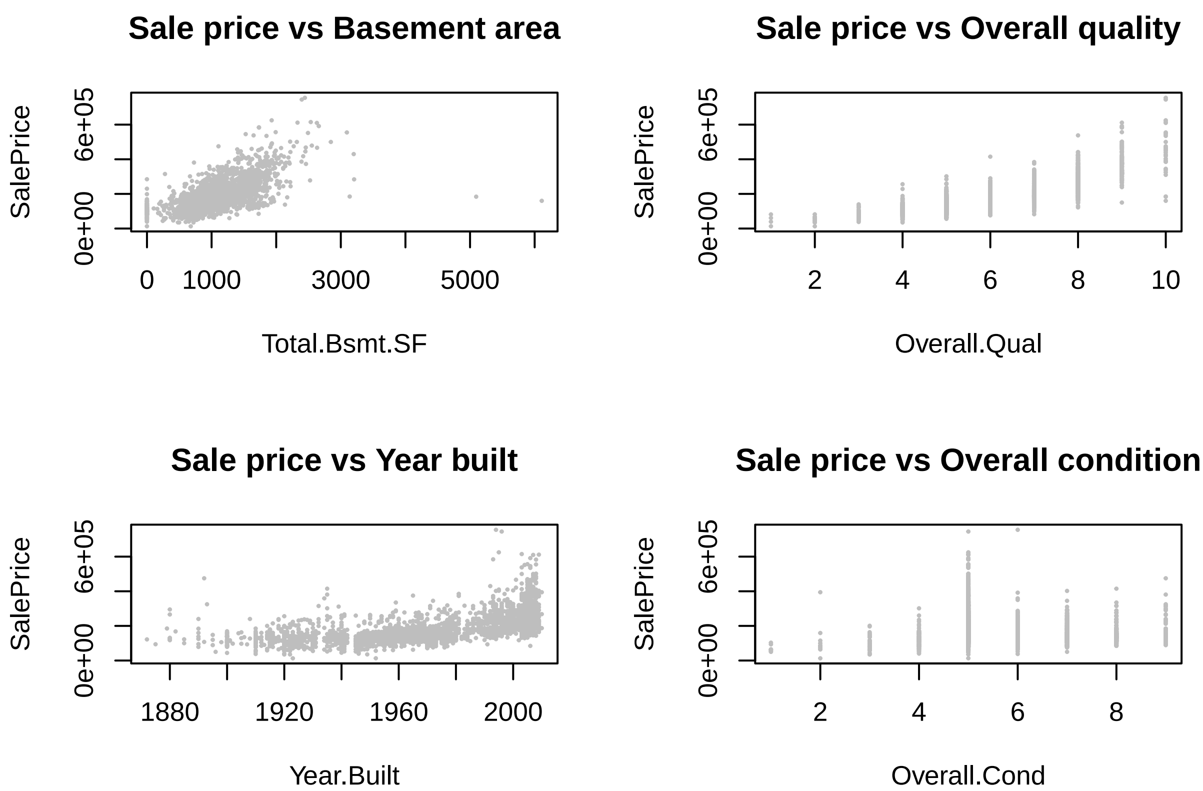
For this section, we'll begin by visualizing the relationships between the house sale prices and some of our predictors. By doing this, we're looking to gain some insight into potential variable transformations we could implement in order to improve the performance of our models.

```
par(mfrow = c(2, 2))

plot(SalePrice ~ Lot.Area, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs Lot area")

plot(SalePrice ~ X1st.Flr.SF, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs First floor area")

plot(SalePrice ~ X2nd.Flr.SF, data = housing_data, col = "grey", pch = 20, cex = 0.3,
```

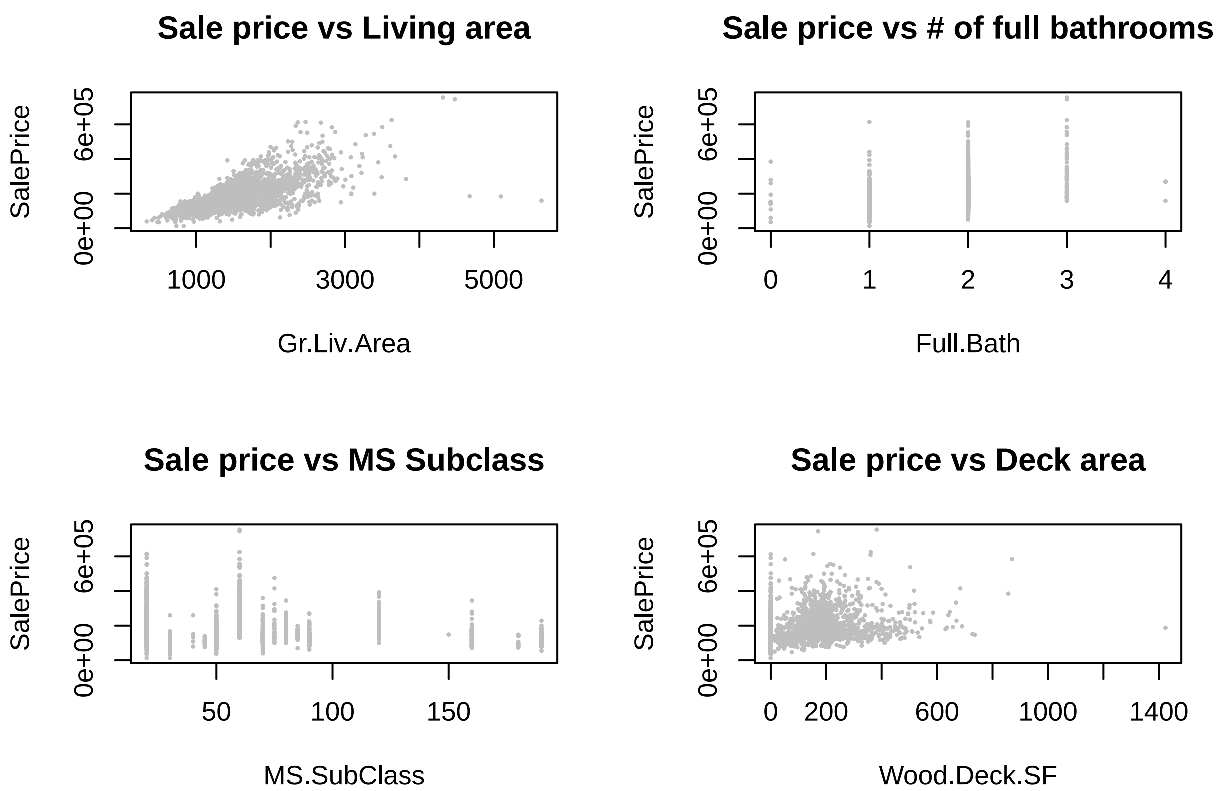
```
     main = "Sale price vs Second floor area")

plot(SalePrice ~ Garage.Area, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs Garage area")
```



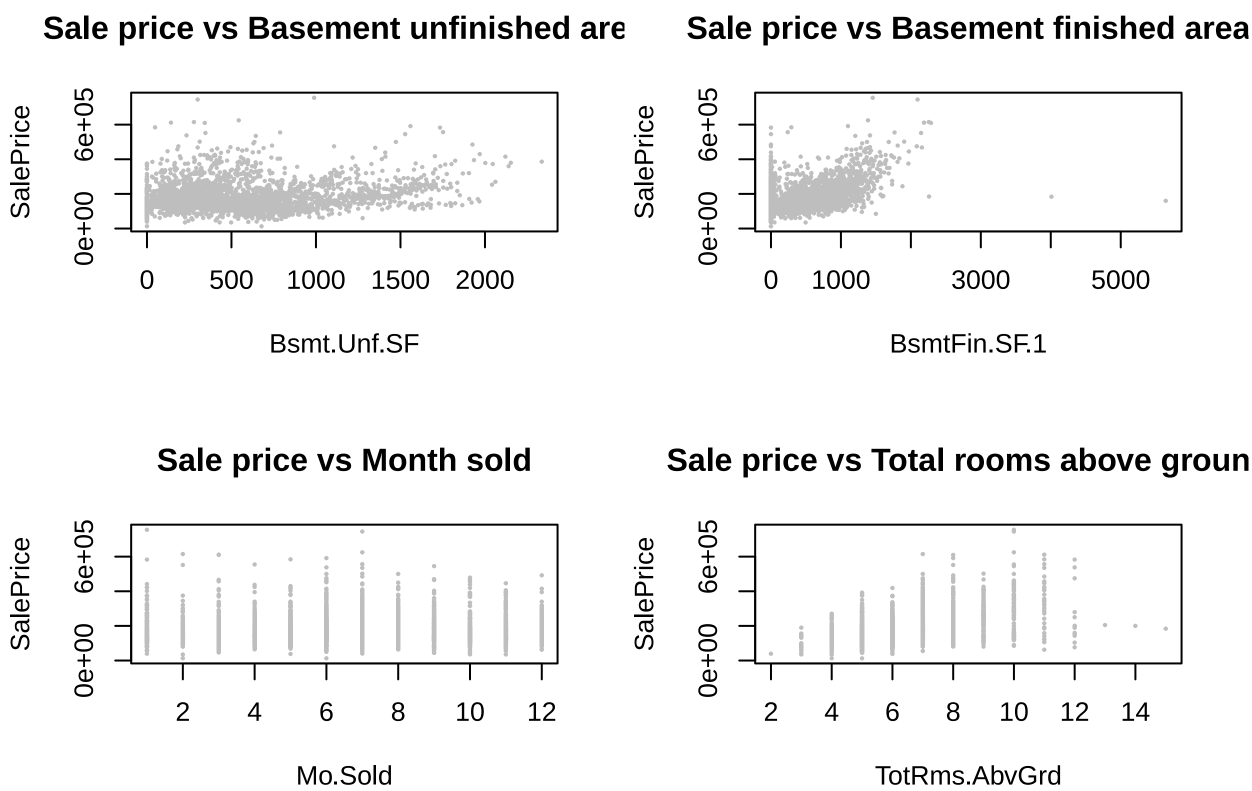

```
par(mfrow = c(2, 2))

plot(SalePrice ~ Total.Bsmt.SF, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs Basement area")

plot(SalePrice ~ Overall.Qual, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs Overall quality")

plot(SalePrice ~ Year.Built, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs Year built")

plot(SalePrice ~ Overall.Cond, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs Overall condition")
```



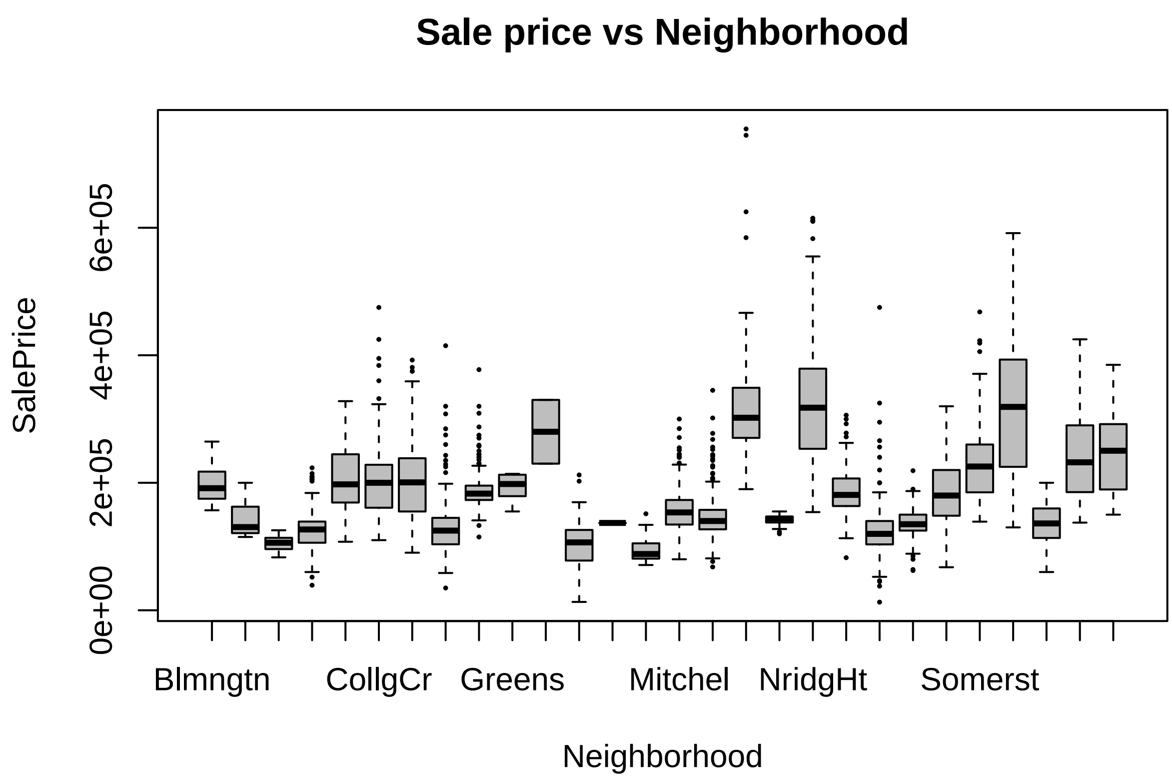

```r
par(mfrow = c(2, 2))

plot(SalePrice ~ Gr.Liv.Area, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs Living area")

plot(SalePrice ~ Full.Bath, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs # of full bathrooms")

plot(SalePrice ~ MS.SubClass, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs MS Subclass")

plot(SalePrice ~ Wood.Deck.SF, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs Deck area")
```




```
par(mfrow = c(2, 2))

plot(SalePrice ~ Bsmt.Unf.SF, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs Basement unfinished area")

plot(SalePrice ~ BsmtFin.SF.1, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs Basement finished area")

plot(SalePrice ~ Mo.Sold, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs Month sold")

plot(SalePrice ~ TotRms.AbvGrd, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs Total rooms above ground")
```




```
plot(SalePrice ~ Neighborhood, data = housing_data, col = "grey", pch = 20, cex = 0.3,
     main = "Sale price vs Neighborhood")
```




```
transformed_model = lm(SalePrice ~ MS.Zoning + log(Lot.Area) + Street + Lot.Shape + Land.Slope + House.S

get_loocv_rmse(saleprice_selected_reduced_backward_aic)
```

```
## [1] 35945.44
```

```
get_loocv_rmse(transformed_model)
```

```
## [1] 31898.11
```

```
sqrt(mean((housing_data$SalePrice - fitted(saleprice_selected_reduced_backward_aic)) ^ 2))
```

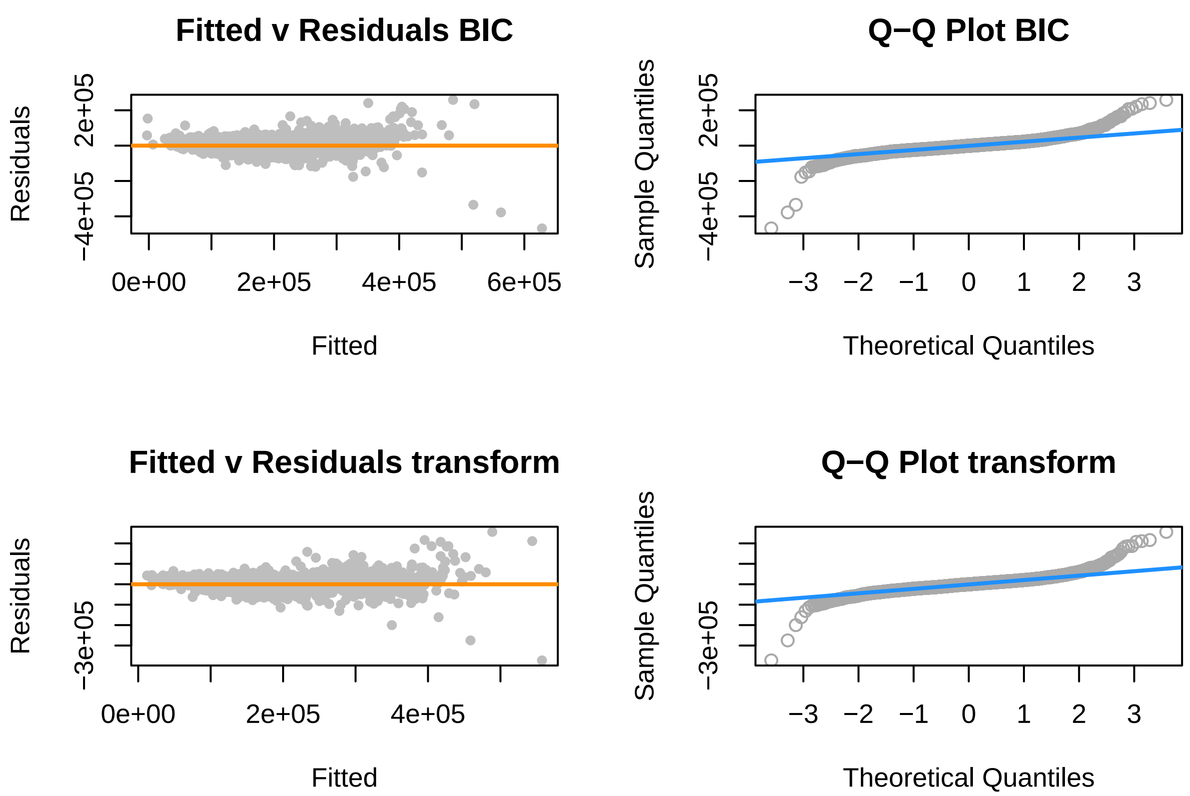
```
## [1] 34069.63
```

```
sqrt(mean((housing_data$SalePrice - fitted(transformed_model)) ^ 2))
```

```
## [1] 29963.36
```

```
par(mfrow = c(2,2))

plot(fitted(saleprice_selected_reduced_backward_aic), resid(saleprice_selected_reduced_backward_aic), co
     xlab = "Fitted", ylab = "Residuals", main = "Fitted v Residuals BIC")
```

```r
abline(h = 0, col = "darkorange", lwd = 2)

qqnorm(resid(saleprice_selected_reduced_backward_aic), main = "Q-Q Plot BIC", col = "darkgrey")
qqline(resid(saleprice_selected_reduced_backward_aic), col = "dodgerblue", lwd = 2)

plot(fitted(transformed_model), resid(transformed_model), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Fitted v Residuals transform")
abline(h = 0, col = "darkorange", lwd = 2)

qqnorm(resid(transformed_model), main = "Q-Q Plot transform", col = "darkgrey")
qqline(resid(transformed_model), col = "dodgerblue", lwd = 2)
```



After adding some polynomial and logarithmic kernel transformations to applicable predictor variables, we can see some noticeable improvement in both RMSE and LOOCV-RMSE metrics when comparing to the reduced backward-BIC model from earlier.