STAT 512 Project

Group 8: Diana Rashidan, Hope Cullers, Garrett Mulcahy, Logan Bradley-Trietsch

Professor Chen

6 December 2019

Introduction

We were tasked to perform statistical analyses of real-world data by applying the knowledge obtained from our STAT512: Applied Regression Analysis course. Our group worked with the Ames Housing data obtained through a website called Kaggle ("House Prices: Advanced Regression Techniques", n.d.) but originally compiled by De Cock (2011). The dataset is a modern and extended version of the more frequently cited Boston Housing dataset compiled in the 70s. We concluded that the real estate data based in Ames, Iowa was simple enough to model and was an intriguing dataset with the potential of meaningful real-world applications. We used the training set of the Ames Housing data which has a large number of observations, specifically, 1460 observations. It has a total of 80 variables which consist of a good mix of 23 nominal, 23 ordinal, 14 discrete and 20 continuous variables. This dataset is also based on individual residential property sales from 2006 to 2010. However, although the dataset is relatively recent, an important limitation in studying this dataset is that inflation and the 2008 housing market crash that happened within this timeframe may act as constraints in predicting current housing prices, even in Ames, Iowa.

In this statistical analysis paper, we want to discover which of the 79 potential explanatory variables have the most significant effect on housing prices. We developed a model of house pricing through a few stages: preliminary predictor screening, selection of appropriate transformations, variable selection, analysis of potential model, identification of outliers, re-analyzing the best model, and interpretation of the final model, which will be discussed further in the "Methods" section. Ultimately, we were able to narrow down the observations and variables to 1448 and 18, respectively. The final model we constructed is `log(SalePrice) ~ OverallQual + log(GrLivArea) + Neighborhood + MSSubClass + OverallCond + log(LotArea) + HasBasement + YearRemodAdd + HasGarage + Fireplaces + MSZoning + KitchenAbvGr + Functional + HasX2ndFlr + SaleCondition + KitchenQual + CentralAir + BedroomAbvGr`.

The ultimate goal of our analysis is to then study the individual relationship between the variables in our model and determine which variable in the dataset, if any, played a significant role in forecasting house prices. We would also like to study the relationship that these significant variables have with housing prices. Such an analysis could potentially help establish an elementary guideline to solve a problem or question we might have related to pricing homes.

Methods

a) Description of data

In this section, we describe the stages of the analysis and justify our decisions. Our data consist of seventy-nine potential explanatory variables with 1460 cases. Our sampling unit is a single house located in Ames, Iowa. The data was from a *Journal of Statistics Education* (De Cock) article that was made publicly available on Kaggle (House). First, we have the continuous predictors, e.g., `LotArea` (lot size in square feet), `LotFrontage` (linear feet of property connected to a street), `MasVnrArea` (masonry veneer area in square feet), `BsmtFinSF1` and `BsmtFinSF2` (type one and type two of square feet of basement area, respectively), etc. Next, we have factors with different levels, e.g., `Neighborhood`. There were also discrete numerical variables with low ranges, e.g., `Bedroom` and `Kitchen`. To keep the discussion of the seventy-nine potential predictors brief, we include the full data dictionary at the beginning of Appendix B. Lastly, our response variable is `SalePrice`, the selling price of the house in USD.

b) Exploratory Analysis

By inspection, we first identified variables with a large numbers of NAs or that were factors with only one level. These variables were removed because large numbers of NAs would interfere with our analysis and factors with only one level do not provide any additional information about the data. For these reasons, we removed `Street` (one level), `Alley` (94% NA), `Utilities` (one level), `LowQualFinSF` (mostly composed of zeros), `FireplaceQu` (47% NA), `Fence` (81% NA) and `MiscFeature` (96% NA). We also removed `MiscVal` because the nature of the data itself limits our ability to make a meaningful interpretation in our regression. Lastly, we dropped `YearBuilt` (original construction date) and kept `YearRemodAdd` since the latter captures both the most recent remodel date and original construction date (if there has been no remodel).

We noticed that there were several subsets of variables that were highly correlated with each other. These subsets included variables that recorded some aspect of a specific attribute of the house, such as the basement, garage, outdoor seating, or a pool. We realized that if the house did not have a basement (or garage, outdoor seating, pool, etc.) then several of the variables in each subset would be either 0 or NA. Thus, in order to reduce the number of highly correlated variables while still retaining useful information (such as the presence of a basement, garage, outdoor seating, pool, etc.) we decided to replace each subset with a single variable. For a group of related predictors, a "new" indicator variable was created as a factor with two levels, either having a certain quality, encoded as 1, or not having the quality, encoded as 0. They are named `HasX` where `X` is the name of the common feature shared by the variables. By defining this new variable, we avoid redundancy in the information provided by the related predictors. Our results are summarized in the table below.

Table 2.1: New Variable Table

Variables	Values
Original: `LotFrontage` **New**: `HasLotFrontage`	0 - has zero linear square feet of lot frontage 1 - has greater than zero linear square feet of lot frontage
Original: `MasVnrArea` and `MasVnrType` **New**: `HasMasVnr`	0 - has zero square feet of masonry veneer area or a `MasVnrType` of "None" 1- has greater than zero square feet of masonry veneer area and a `MasVnrType` not equal to "None"
Original: `BsmtQual`, `BsmtCond`, `BsmtExposure`, `BsmtFinType1`, `BsmtFinSF1`, `BsmtFinType2`, `BsmtFinSF2`, `BsmtUnfSF`, and `TotalBsmtSF` **New**: `HasBasement`	0 - has zero square feet of basement 1 - has greater than zero square feet of basement
Original: `X2ndFlrSF` **New**: `HasX2ndFlr`	0 - has zero second floor square feet 1 - has greater than zero second floor square feet
Original: `GarageType`, `GarageYrBlt`, `GarageFinish`, `GarageCars`, `GarageArea`, `GarageQual` and `GarageCond` **New**: `HasGarage`	0 - has zero square feet of garage 1 - has greater than zero square feet of garage
Original: `WoodDeckSF`, `OpenPorchSF`, `EnclosedPorch`, `X3SsnPorch`, and `ScreenPorch` **New**: `HasOutdoorSeating`	0 - has zero square feet of any of the five specified predictors 1 - has greater than zero square feet of any of the five specified predictors
Original: `PoolArea` and `PoolQC` **New**: `HasPool`	0 - has zero square feet of pool 1 - has greater than zero square feet of pool

Lastly, it is important to note that the variables MSSubClass, OverallQual, OverallCond and MoSold are factors even though their factor levels are numerical. We conclude our preliminary predictor screening by compiling a "cleaned" dataset that we continued to work with throughout our analysis.

Next, we examined the response variable SalePrice and the remaining continuous predictors. Predictors that represent the number of bathrooms, bedrooms, and kitchens had less than one order of magnitude and so were not considered for transformation since any transformation is unlikely to be helpful. This fact is by the Range Rule discussed in Chapter 8 (Weisberg). We also do not consider YearRemodAdd for transformation since the variable is badly scaled. As for SalePrice, LotArea, X1stFlrSF, and GrLivArea, the univariate distributions show strong positive skew and range over at least one order of magnitude. The bivariate relationships, while positive, definitely display nonconstant variance, as shown in the scatterplot matrix Figure 1 in Appendix B (all Figure references will be found in Appendix B). From the Log Rule in Chapter 8 and the results of our power transformation analysis (Figure 2), a log transformation for each of the predictors was deemed appropriate. We verified these results by looking at marginal tests of transformations for each predictor with the response (Figures 3-5), and these tests also suggested the log transform. According to Weisburg Chapter 7, a log transformation may be appropriate for the response to stabilize variance. To confirm this intuition, we employed the Inverse Response Plot and Box-Cox methods. Both methods suggested that the log transform was appropriate (Figures 6-8). The scatterplot matrix of the log transformed predictors and log transformed response variable in Figure 10 shows that the bivariate relationships are all positive, linear, and possess constant variance. Finally, we performed a correlation test between each transformed predictor versus the transformed response. The results returned positive correlations that were all significant due to their very low p-values (Figure 9).

c) Model Building Process

By this stage, we had a total of fifty-two variables, where forty were factors and twelve were numeric. During variable selection, we ran 1) AIC backward elimination, 2) AIC forward selection, 3) BIC backward elimination, and 4) BIC forward selection. We did not receive any conclusive output from the first three, but succeeded to select twenty variables of interest via BIC forward selection (Figure 11). We then used these twenty variables to fit a model.

d) Model Diagnostics

The fitted model was tested for outliers (Figure 12), but we decided not to remove any cases since there were few outliers relative to the size of the dataset. Next, we checked for influential cases, which we considered to be cases with leverage equal to one (and thus a non-defined Cook's Distance) or cases with much higher Cook's Distance relative to all other cases. In Figure 13, we verify that the influential points were highly separated from others, making their removal straightforward. Figure 14 shows cases with NaN for Cook's Distance, indicating that their leverage is one and hence Cook's Distance cannot be calculated. After removing the influential points, we refit the model and check for outliers again. This time, we found no influential points and proceeded to tweak our model according to our findings.

We refit the model with the above-mentioned influential points removed. The ANOVA table for this model (Figure 17) discovered that `Condition2` and `RoofMatl` were potentially redundant or did not add any additional information (since they were no longer significant). To test this, we performed two partial F-tests. First, we tested to see if the fit of the model was significantly different with `RoofMatl` excluded. The large p-value from Figure 18 shows that it does not, so we decided to remove the variable from the model. The ANOVA table in Figure 19 for the model with `RoofMatl` removed shows that `Condition2` is still not significant. Thus, we perform another partial F-test on this reduced model to see if the fit of the model is significantly different with `Condition2` excluded. Again, the large p-value from Figure 20 leads us to conclude that the fit is not significantly different, so we decided to remove `Condition2` from the model. The ANOVA table for this newest mean function (`Condition2` and `RoofMatl` removed) shows that all terms are significant (this ANOVA table is in the Results section).

Next, we examined the residual plots to test for curvature and appropriateness of the mean function (Figures 21-24). Only `log(GrLivArea)` showed a significant curvature in the plot as shown in Figure 21 and the test in Figure 24. We considered adding a quadratic or cubic term to the model to reduce the curvature, but we decided not to since transforming an already transformed predictor is unusual. We acknowledge this shortcoming in our model which suggests that our mean function may not be optimal.

When we tested for outliers again on the reduced model, there were few outliers and no points with extremely large influence (Figures 25 and 26), so we decided to not exclude any more points from our analysis. Finally, we ran some diagnostic methods to check model assumptions. We checked for nonconstant variance which returned very strong evidence against the assumption of constant variance as shown in Figure 28. This makes sense given a large number of predictors. To fix this, we use the hccm sandwich estimator for variance. We decided to use a sandwich estimator for variance as opposed to finding explicit weights because we have no prior reason to believe that any one of the eighteen variables in the model should have such a direct relationship with variance. We also check the normality assumption and observed in Figure 27 that the residuals do not appear to be normally distributed. We acknowledge this flaw, but since our sample size is very large and the regression procedure is robust toward violations of normality, we do not dwell too much on this shortcoming. Thus, all of the diagnostics have been checked.

e) **Description of Inferential Methods**

Our inferential methods involve using an ANOVA table to screen for potentially redundant predictors and to ensure that all of our predictors were significant at the 0.05 significance level, although we would like to note that all predictors except three are significant at the 0.001 significance level. To study the relationship that these predictors have with `log(SalePrice)` we employ two strategies depending on the variable type. First, for continuous predictors and factors with two levels, we looked at the coefficient estimates and their corresponding tests to determine the strength and direction of the association. For factors with multiple levels, we examine several effects plots to determine the differences in `log(SalePrice)` between the levels.

Results

We ultimately considered the following two models. While Model 1 was obtained through variable selection, we decided to work with Model 2 based on two partial F-tests giving us p-values of 0.3237 and 0.1102 respectively, showing that the variables `RoofMatl` and `Condition2` did not significantly improve the fit of the model.

Model Selection Table

Model 1	Model 2 (best model)
log(SalePrice) ~ OverallQual + log(GrLivArea) + Neighborhood + MSSubClass + OverallCond + log(LotArea) + HasBasement + YearRemodAdd + HasGarage + Fireplaces + MSZoning + KitchenAbvGr + Functional + HasX2ndFlr + SaleCondition + KitchenQual + CentralAir + BedroomAbvGr + RoofMatl + Condition2	log(SalePrice) ~ OverallQual + log(GrLivArea) + Neighborhood + MSSubClass + OverallCond + log(LotArea) + HasBasement + YearRemodAdd + HasGarage + Fireplaces + MSZoning + KitchenAbvGr + Functional + HasX2ndFlr + SaleCondition + KitchenQual + CentralAir + BedroomAbvGr

To correct for nonconstant variance, we use a sandwich estimator to obtain the standard error of the coefficients. With this correction, we display the ANOVA table and table of coefficient estimates, along with the t statistic, p-value, and 95% confidence interval for each estimate.

Table 3.1: ANOVA Table for Model

```
## Coefficient covariances computed by hccm
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: log(SalePrice)
##                  Df        F    Pr(>F)
## OverallQual       8  22.4772 < 2.2e-16 ***
## log(GrLivArea)    1 575.6909 < 2.2e-16 ***
## Neighborhood     24   7.6016 < 2.2e-16 ***
## MSSubClass       14   5.6591 8.741e-11 ***
## OverallCond       7   7.7303 3.440e-09 ***
## log(LotArea)      1  53.0067 5.592e-13 ***
## HasBasement       1  25.4595 5.126e-07 ***
## YearRemodAdd      1  20.2354 7.428e-06 ***
## HasGarage         1  16.5373 5.043e-05 ***
## Fireplaces        1  17.7507 2.683e-05 ***
## MSZoning          4   2.7980  0.024879 *
## KitchenAbvGr      1  16.3997 5.417e-05 ***
## Functional        5   8.1366 1.404e-07 ***
## HasX2ndFlr        1  15.8741 7.127e-05 ***
## SaleCondition     5   5.8738 2.227e-05 ***
## KitchenQual       3   7.6297 4.646e-05 ***
```

```
## CentralAir        1   6.8913  0.008758 **
## BedroomAbvGr      1   6.3087  0.012130 *
## Residuals      1366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 3.2: Parameter Coefficients for Model

```
##
## t test of coefficients:
##
##                        Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)           3.44977318  0.63970167  5.3928 8.166e-08 ***
## OverallQual3          0.01868797  0.09254828  0.2019 0.8400041
## OverallQual4          0.06667956  0.08034642  0.8299 0.4067398
## OverallQual5          0.08909006  0.08071980  1.1037 0.2699198
## OverallQual6          0.12886145  0.08001344  1.6105 0.1075203
## OverallQual7          0.19692842  0.08104049  2.4300 0.0152269 *
## OverallQual8          0.29323922  0.08214911  3.5696 0.0003699 ***
## OverallQual9          0.42930219  0.08645512  4.9656 7.712e-07 ***
## OverallQual10         0.55252728  0.09340938  5.9151 4.186e-09 ***
## log(GrLivArea)        0.58745052  0.02448367 23.9936 < 2.2e-16 ***
## NeighborhoodBlueste   0.05124736  0.05107322  1.0034 0.3158410
## NeighborhoodBrDale   -0.01847935  0.04732769 -0.3905 0.6962608
## NeighborhoodBrkSide   0.01135369  0.04436045  0.2559 0.7980343
## NeighborhoodClearCr   0.04468239  0.05046808  0.8854 0.3761186
## NeighborhoodCollgCr   0.02949559  0.03113365  0.9474 0.3436096
## NeighborhoodCrawfor   0.08711203  0.04046389  2.1528 0.0315064 *
## NeighborhoodEdwards  -0.06813644  0.03722755 -1.8303 0.0674275 .
## NeighborhoodGilbert  -0.00692616  0.03245797 -0.2134 0.8310558
## NeighborhoodIDOTRR   -0.07201596  0.05170783 -1.3927 0.1639227
## NeighborhoodMeadowV  -0.10660979  0.04670469 -2.2826 0.0226050 *
## NeighborhoodMitchel  -0.00225492  0.03528021 -0.0639 0.9490475
## NeighborhoodNAmes    -0.04262373  0.03354964 -1.2705 0.2041344
## NeighborhoodNoRidge   0.12584749  0.03567162  3.5279 0.0004327 ***
## NeighborhoodNPkVill   0.05091689  0.03975111  1.2809 0.2004490
## NeighborhoodNridgHt   0.08511133  0.03086063  2.7579 0.0058945 **
## NeighborhoodNWAmes   -0.04882184  0.03494274 -1.3972 0.1625817
## NeighborhoodOldTown  -0.10137960  0.04556872 -2.2248 0.0262600 *
## NeighborhoodSawyer   -0.05808258  0.03624006 -1.6027 0.1092282
## NeighborhoodSawyerW  -0.00604010  0.03379094 -0.1787 0.8581611
## NeighborhoodSomerst   0.03732297  0.05055467  0.7383 0.4604775
## NeighborhoodStoneBr   0.12884402  0.03921166  3.2859 0.0010425 **
## NeighborhoodSWISU    -0.02724526  0.04430395 -0.6150 0.5386820
## NeighborhoodTimber    0.02561440  0.03475851  0.7369 0.4612947
## NeighborhoodVeenker   0.06106258  0.04057532  1.5049 0.1325760
## MSSubClass30         -0.11925885  0.03003244 -3.9710 7.530e-05 ***
## MSSubClass40         -0.12255603  0.15993199 -0.7663 0.4436297
## MSSubClass45         -0.08227461  0.03449747 -2.3849 0.0172174 *
```

```
## MSSubClass50         -0.06437302  0.02981943 -2.1588 0.0310424 *
## MSSubClass60          0.02181186  0.02571058  0.8484 0.3963854
## MSSubClass70         -0.10106533  0.03381396 -2.9889 0.0028502 **
## MSSubClass75         -0.10942935  0.03867514 -2.8294 0.0047310 **
## MSSubClass80          0.00945539  0.01623818  0.5823 0.5604648
## MSSubClass85          0.07376573  0.01671337  4.4136 1.097e-05 ***
## MSSubClass90          0.03563258  0.03377842  1.0549 0.2916614
## MSSubClass120         0.01255586  0.02262610  0.5549 0.5790347
## MSSubClass160        -0.02524725  0.03459061 -0.7299 0.4655842
## MSSubClass180         0.10489217  0.04034548  2.5998 0.0094270 **
## MSSubClass190        -0.00573273  0.03291379 -0.1742 0.8617544
## OverallCond3         -0.09493307  0.09629260 -0.9859 0.3243660
## OverallCond4          0.00367800  0.08854344  0.0415 0.9668723
## OverallCond5          0.05564168  0.08593158  0.6475 0.5174098
## OverallCond6          0.08275418  0.08659406  0.9557 0.3394149
## OverallCond7          0.11673427  0.08660474  1.3479 0.1779150
## OverallCond8          0.11388177  0.08858400  1.2856 0.1988078
## OverallCond9          0.14962201  0.08984875  1.6653 0.0960892 .
## log(LotArea)          0.08980798  0.01233530  7.2806 5.592e-13 ***
## HasBasement1          0.15238567  0.03020087  5.0457 5.126e-07 ***
## YearRemodAdd          0.00138814  0.00030859  4.4984 7.428e-06 ***
## HasGarage1            0.09612831  0.02363847  4.0666 5.043e-05 ***
## Fireplaces            0.02797734  0.00664048  4.2132 2.683e-05 ***
## MSZoningFV            0.40490693  0.12217843  3.3141 0.0009436 ***
## MSZoningRH            0.36499968  0.12315490  2.9637 0.0030918 **
## MSZoningRL            0.35246623  0.11499726  3.0650 0.0022194 **
## MSZoningRM            0.33671400  0.11404407  2.9525 0.0032061 **
## KitchenAbvGr         -0.11643882  0.02875275 -4.0497 5.417e-05 ***
## FunctionalMaj2       -0.22599245  0.08455082 -2.6729 0.0076103 **
## FunctionalMin1        0.00507519  0.04652029  0.1091 0.9131421
## FunctionalMin2        0.01480525  0.04516711  0.3278 0.7431221
## FunctionalMod         0.01640709  0.08074629  0.2032 0.8390143
## FunctionalTyp         0.09135508  0.04013796  2.2760 0.0229986 *
## HasX2ndFlr1          -0.10125403  0.02541369 -3.9842 7.127e-05 ***
## SaleConditionAdjLand  0.11101529  0.08961756  1.2388 0.2156446
## SaleConditionAlloca   0.08644498  0.07799173  1.1084 0.2678901
## SaleConditionFamily  -0.00230347  0.05373399 -0.0429 0.9658131
## SaleConditionNormal   0.07324852  0.01980032  3.6994 0.0002247 ***
## SaleConditionPartial  0.13073817  0.02517512  5.1931 2.381e-07 ***
## KitchenQualFa        -0.13357960  0.03574303 -3.7372 0.0001937 ***
## KitchenQualGd        -0.06106139  0.02069855 -2.9500 0.0032315 **
## KitchenQualTA        -0.09246873  0.02258895 -4.0935 4.497e-05 ***
## CentralAirY           0.06155304  0.02344766  2.6251 0.0087584 **
## BedroomAbvGr         -0.01964327  0.00782070 -2.5117 0.0121296 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 3.3: 95% Confidence Intervals for Parameter Estimates

```
##                            2.5 %       97.5 %
## (Intercept)           2.1948690363  4.704677329
## OverallQual3         -0.1628641848  0.200240126
## OverallQual4         -0.0909361780  0.224295304
## OverallQual5         -0.0692581570  0.247438273
## OverallQual6         -0.0281010816  0.285823980
## OverallQual7          0.0379511261  0.355905716
## OverallQual8          0.1320871394  0.454391310
## OverallQual9          0.2597029945  0.598901387
## OverallQual10         0.3692859054  0.735768653
## log(GrLivArea)        0.5394208374  0.635480193
## NeighborhoodBlueste  -0.0489430937  0.151437812
## NeighborhoodBrDale   -0.1113221777  0.074363485
## NeighborhoodBrkSide  -0.0756682899  0.098375676
## NeighborhoodClearCr  -0.0543209427  0.143685731
## NeighborhoodCollgCr  -0.0315793626  0.090570546
## NeighborhoodCrawfor   0.0077339349  0.166490121
## NeighborhoodEdwards  -0.1411657967  0.004892919
## NeighborhoodGilbert  -0.0705990303  0.056746713
## NeighborhoodIDOTRR   -0.1734513156  0.029419393
## NeighborhoodMeadowV  -0.1982304909 -0.014989096
## NeighborhoodMitchel  -0.0714641954  0.066954348
## NeighborhoodNAmes    -0.1084381291  0.023190667
## NeighborhoodNoRidge   0.0558703959  0.195824591
## NeighborhoodNPkVill  -0.0270629581  0.128896732
## NeighborhoodNridgHt   0.0245719642  0.145650700
## NeighborhoodNWAmes   -0.1173690971  0.019725409
## NeighborhoodOldTown  -0.1907718461 -0.011987345
## NeighborhoodSawyer   -0.1291747831  0.013009618
## NeighborhoodSawyerW  -0.0723278577  0.060247649
## NeighborhoodSomerst  -0.0618502401  0.136496180
## NeighborhoodStoneBr   0.0519224283  0.205765609
## NeighborhoodSWISU    -0.1141564114  0.059665889
## NeighborhoodTimber   -0.0425714399  0.093800245
## NeighborhoodVeenker  -0.0185341143  0.140659283
## MSSubClass30         -0.1781735419 -0.060344155
## MSSubClass40         -0.4362949536  0.191182895
## MSSubClass45         -0.1499483824 -0.014600843
## MSSubClass50         -0.1228698581 -0.005876184
## MSSubClass60         -0.0286246366  0.072248351
## MSSubClass70         -0.1673982403 -0.034732412
## MSSubClass75         -0.1852984557 -0.033560253
## MSSubClass80         -0.0223990784  0.041309865
## MSSubClass85          0.0409790692  0.106552390
## MSSubClass90         -0.0306306283  0.101895782
## MSSubClass120        -0.0318298098  0.056941527
## MSSubClass160        -0.0931037166  0.042609224
## MSSubClass180         0.0257463573  0.184037973
## MSSubClass190        -0.0702997866  0.058834325
## OverallCond3         -0.2838304753  0.093964344
## OverallCond4         -0.1700178508  0.177373852
## OverallCond5         -0.1129304942  0.224213856
## OverallCond6         -0.0871175692  0.252625921
## OverallCond7         -0.0531584436  0.286626980
```

```
## OverallCond8         -0.0598936557  0.287657201
## OverallCond9         -0.0266344885  0.325878503
## log(LotArea)          0.0656098017  0.114006166
## HasBasement1          0.0931405502  0.211630781
## YearRemodAdd          0.0007827853  0.001993499
## HasGarage1            0.0497566762  0.142499954
## Fireplaces            0.0149507044  0.041003976
## MSZoningFV            0.1652292417  0.644584624
## MSZoningRH            0.1234064557  0.606592908
## MSZoningRL            0.1268758636  0.578056595
## MSZoningRM            0.1129935070  0.560434490
## KitchenAbvGr         -0.1728431497 -0.060034500
## FunctionalMaj2       -0.3918559766 -0.060128933
## FunctionalMin1       -0.0861837670  0.096334151
## FunctionalMin2       -0.0737991795  0.103409673
## FunctionalMod        -0.1419930877  0.174807273
## FunctionalTyp         0.0126163528  0.170093800
## HasX2ndFlr1          -0.1511081306 -0.051399938
## SaleConditionAdjLand -0.0647876750  0.286818255
## SaleConditionAlloca  -0.0665515763  0.239441531
## SaleConditionFamily  -0.1077135421  0.103106609
## SaleConditionNormal   0.0344061907  0.112090840
## SaleConditionPartial  0.0813520797  0.180124262
## KitchenQualFa        -0.2036967924 -0.063462416
## KitchenQualGd        -0.1016657684 -0.020457002
## KitchenQualTA        -0.1367815205 -0.048155930
## CentralAirY           0.0155557114  0.107550373
## BedroomAbvGr         -0.0349851498 -0.004301388
```

Forward variable selection with BIC as criterion selected twenty variables of interest. After removing outliers and influential cases, we were left with eighteen variables of interest, those given in Model 2 above. The ANOVA table confirms that each of these terms is significant in the model (at a significance level of 0.05, with all but three significant at the 0.001 significance level). Thus, we conclude that each of the 18 variables in the model is important in determining the sale price of a house in Ames, Iowa.

Now, we wish to investigate the effect of these eighteen variables on the sale price. First, we examine the effect of numerical predictors on the response, which can be quickly determined from the sign of the estimate. With all other regressors assumed constant, each of GrLivArea, LotArea, YearRemodAdd, and Fireplaces have a positive relationship with `log(SalePrice)`. Since `LotArea` and `GrLivArea` are both in log scale, differences in `log(SalePrice)` are most pronounced when `LotArea` and `GrLivArea` are both small; that is, increasing an already large lot is not expected to increase `log(SalePrice)` that much. The positive relationship with `YearRemodAdd` suggests that more recently remodeled houses have a higher sales price, and similarly, we expect that the more fireplaces a house has the larger its sales price will be. Interestingly, with all other regressors assumed constant, each of `KitchenAbvGr` and `BedroomAbvGr` has a negative relationship with `log(SalePrice)`. These relationships are certainly unexpected - surprisingly, having more bedrooms or kitchens in a house would decrease the sale price; we attribute this unexpected relationship to the presence of other terms in the model which may already encode some of this information.

For two-level factors, the sign of the estimate of the dummy variable used to include the factor in the model is also indicative of the relationship between the factor and `log(SalePrice)`. When all other regressors are assumed constant, we expect that having a basement, having a garage, and having central air conditioning will increase `log(SalePrice)`, but there is the unusual relationship that having a second floor is expected to lower sale price. Again, we attribute this unusual behavior to the presence of other terms in the model which may already encode some of this information, such as `MSSubClass`, in which a building with a second story must have a code of 60, 70, 75, 80, or 160.

To interpret the effect of the factors with several levels on the response, we examine the effects plots corresponding to these factors. We include these additional plots in Appendix B. From Figure 29, it appears that as the `Overall Quality` rating increases, the fitted mean `log(SalePrice)` increases as well. However, with the 95% confidence intervals, we see that the difference in fitted mean `log(SalePrice)` for quality ratings (in the 2-5 range) is not significantly different from each other. However, high-quality ratings (in the 8-10 range) do appear to have a fitted mean `log(SalePrice)` that is larger than that associated with lower quality ratings. Similarly, it appears that as the `Overall Condition` (Figure 32) rating increases the fitted mean `log(SalePrice)` increases as well. However, the confidence intervals are much wider than for the previous rating, so it appears that this rating has a weaker relationship with sale price than `Overall Quality`.

Concerning `Kitchen Quality` (Figure 36), it appears that an "Excellent" rating is associated with a higher sale price, while the differences between the other three ratings are not significant. Similarly, for `MSZoning` (Figure 33), a classification of "Commercial" is associated with a lower sale price, while the differences between all other classifications are not significant. For the factor `Functional` (home functionality), a rating of Major Deductions 2 is associated with a lower sale price, while the differences between Major Deductions 1, Minor Deductions 1 and 2, and Moderate Deductions is not significant (Figure 34). While the difference between Typical and Moderate deductions is not significant, a Typical rating has a significantly larger sale price than all other ratings excluding Moderate. For `SaleCondition`, a Partial rating (meaning that the home was likely a new construction) is associated with a larger sale price than that of Normal, Abnormal, and Family sales (Figure 35). However, the fitted mean sale price for a partial rating is not significantly different than that associated with Adjoining Land Purchases or Allocation.

The factors `Neighborhood` and `MsSubClass` (type of dwelling sold) exhibit the most interesting behavior with `log(SalePrice)`, and will be expanded upon in the Discussions section. From the effects plot, it is clear that there are significant differences between many of the levels. With regards to `Neighborhood` (Figure 30), we note that the neighborhoods of Northridge and Stone Brook have the largest fitted mean for `log(SalePrice)`, and while these fitted means may not be statistically different from some other neighborhoods, we decided to focus on specific trends in these neighborhoods in the Discussion section. Similarly, Old Town and Meadow Village appear to have the lowest fitted mean for `log(SalePrice)`. For `MSSubClass` (Figure 31), we also see that some levels of the factor have significant behavior while others do not; for example, Multilevel Planned Urban Developments (Class 180) appear to have the largest fitted mean `log(SalePrice)`, while 1-story buildings with finished attic (Class 40) appear to have the greatest variability in fitted mean `log(SalePrice)`.

Discussion

In conclusion, and as expected, many factors affect housing prices. Within the model, we see that the simple presence of conditions such as a garage and central air are associated with higher sale prices. This makes sense practically, as one would expect the inclusion of valued conditions to correlate with increased overall value. Likewise, we see that variables associated with space (`GrLivArea`, `LotArea`), quality and condition (`OverallQual`, `KitchenQual`, `OverallCond`), remodeling (`YearRemodAdd`), and multiple presences of a condition (`Fireplaces`, `KitchenAbvGr`, `BedroomAbvGr`) have a relationship with a house's sale price.

We also see in the model that some categorical conditions, such as neighborhood (`Neighborhood`) and the type of dwelling (`MSSubClass`), have a relationship with a house's sale price. This is an interesting finding because it leads to more questions as to why certain neighborhoods and dwellings have different sale prices than others. For example, as mentioned in Results, Old Town has an especially low fitted mean `log(SalePrice)`. As it happens, the Old Town Historic District is an official listing in the National Register of Historic Places (National Register of Historic Places). According to the Historic Preservation Commission of Ames, you must apply and be approved to make any changes to the exterior of your home if you live in Old Town (Historic Preservation Commission). While more research would be needed to determine whether this fact is truly impacting the house prices within this neighborhood, it is an interesting finding.

Understanding house prices help buyers determine whether a house is worth its asking price and at what price to make an offer. Likewise, this understanding helps sellers determine at what price to list their houses, whether an offer meets the value of a house, and at what price to make a counter-offer. The findings of this report help in defining what parameters affect housing prices and are useful for individuals currently involved in the housing market. While the findings of this report only directly apply to the area within the boundary of Ames, Iowa, this model may apply to other areas of the United States, perhaps other smaller cities in Iowa or other Midwestern states, as well.

This leads to the main limitation of the data, which is that the data only includes sold houses in Ames, Iowa. As such, the model cannot accurately be applied to any other location without further testing. In addition, some other variables not included in the dataset might potentially be useful in this model. Such variables include demographics within each neighborhood, how long the houses were listed before they sold, and how many offers and counter-offers led up to the closing. We consulted with a housing expert who stated that such variables also tend to be very influential.

Within the housing industry, it is well-known that the housing market tends to fluctuate in a semi-cyclical manner. These fluctuations occur as a result of various factors, as described in a paper by Dennis R. Capozza and others. In their research, they observed the pattern of the ever-changing housing market and concluded that "[h]ouse prices react differently to economic shocks depending on such factors as growth rates, area size, and construction costs," (Capozza et. All). There is no way for us to gauge this semi-cyclic fluctuation with the dataset that we have and the model that we built from it.

Lastly, the timeframe of the data is, in and of itself, a limiting factor. The data only includes houses sold between January 2006 and July 2010. Along with this, the economy – and by association, the housing market – saw a major crash in 2008. In a paper discussing this crisis, Reavis states that "the housing boom from the late 1990s into the mid-2000s drove much of the U.S. economy" and that the ratio of household debt to GDP doubled from 50% to 100% from the 1980s to the mid-2000s. (Reavis 3). For context, "[t]he last time the level of debt was 100% of GDP was 1929, the beginning of the Great Depression" (Reavis 4). Interestingly, and despite this information, the variable associated with the year sold (`YrSold`) was not selected during the variable selection process. Knowing that there was indeed a crash in the market in 2008, this adds to the idea that a larger time frame is needed in order to see the impacts of economic changes and time.

While a great deal of more research into this topic is needed to truly support the housing industry, this model provides certainty that there indeed is a relationship between various house conditions and sale price, as well as provides useful insight into what these correlations might be. Further data involving locations beyond Ames, Iowa as well as a larger time frame would add more confidence to the model in application to the housing industry throughout the United States. Yet even with the limits of the data available, this model would certainly prove valuable to experts in the field.

References and Appendices

Capozza, D. R., Hendershott, P. H., Mack, C., & Mayer, C. J. (2002, October). Determinants of Real House Price Dynamics. *NBER Working Paper Series No. 9262*. Retrieved from https://www.nber.org/digest/may03/w9262.html

De Cock, D. (2011). Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3), 1-15. doi:10.1080/10691898.2011.11889627

Historic Preservation Commission. (n.d.) Retrieved from https://www.cityofames.org/government/departments-divisions-i-z/planning/historic-preservation/historic-preservation-commission

House Prices: Advanced Regression Techniques. (n.d.). Retrieved from https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data.

National Register of Historic Places, Old Town Historic District, Ames, Story County, Iowa, National Register #03001349. Retrieved from https://www.nps.gov/subjects/nationalregister/database-research.htm

Reavis, Cate. (2012, March 16). The Global Financial Crisis of 2008: The Role of Greed, Fear, and Oligarchs. Retrieved from https://mitsloan.mit.edu/LearningEdge/CaseDocs/09-093%20The%20Financial%20Crisis%20of%202008.Rev.pdf

Weisburg, S. (2014) Applied Linear Regression. Hoboken, NJ: John Wiley & Sons, Inc.

Acknowledgements

We would like to thank Vince Cullers, a housing expert, for his consultation on this project.

Appendix A

```r
# Libraries ----
library("alr4")
library("GGally")
library("ggplot2")
library("knitr")
library("lmtest")

# Read in data and discuss data ----
# Data can be downloaded from:
# https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
# Place data in working directory
house <- read.csv("train.csv")
dim(house)
colnames(house)

# Analysis of NAs in data ----
colMeans(is.na(house))

# Create Indicator Variables ----
# X2ndFlr ----
house$HasX2ndFlr <- as.factor(with(house, ifelse(X2ndFlrSF == 0, 0, 1)))
# HasPool ----
house$HasPool <- as.factor(with(house, ifelse(PoolArea == 0, 0, 1)))


# HasBasement ----
house$HasBasement <- as.factor(with(house, ifelse(TotalBsmtSF == 0, 0, 1)))

# HasGarage ----
house$HasGarage <- as.factor(with(house, ifelse(GarageArea == 0, 0, 1)))

# HasOutdoorSeating ----
house$HasOutdoorSeating <- as.factor(
  with(house, ifelse(WoodDeckSF != 0 |
                     OpenPorchSF != 0 |
                     EnclosedPorch != 0 |
                     X3SsnPorch != 0 |
                     ScreenPorch != 0, 1, 0)))
# HasLotFrontage ----
house$HasLotFrontage <- as.factor(with(house, ifelse(LotFrontage == 0 | is.na(LotFrontage), 0, 1)))

# HasMasVnr ----
house$HasMasVnr <- as.factor(with(
  house, ifelse(MasVnrArea == 0 | MasVnrType == "None" |
                is.na(MasVnrArea) | is.na(MasVnrType), 0, 1)))
house$MasVnrType[is.na(house$MasVnrType)] <- "None"

# MoSold ----
house$MoSold <- as.factor(house$MoSold)
```

```r
# Make master file ----
colnames(house)
house <- house[-c(1, 4, 6, 7, 10, 20, 27, 31, 32, 33, 34, 35, 36, 37, 38, 39, 46,
                  58, 59, 60, 61, 62, 63, 64, 65, 67, 68, 69, 70, 71, 72, 73, 74, 75,
                  76)]
# Save as csv ----
write.csv(house, "Housing Data - Master.csv")
# Transformations ----
# Loading in desired data set
house <- read.csv("Housing Data - Master.csv")

# Taking only the desired columns
house <- house[-c(1,30,55,56)]
# Casting categorical variables as factors
house$MSSubClass <- as.factor(house$MSSubClass)
house$OverallQual <- as.factor(house$OverallQual)
house$OverallCond <- as.factor(house$OverallCond)
house$MoSold <- as.factor(house$MoSold)
house$HasX2ndFlr <- as.factor(house$HasX2ndFlr)
house$HasPool <- as.factor(house$HasPool)
house$HasBasement <- as.factor(house$HasBasement)
house$HasGarage <- as.factor(house$HasGarage)
house$HasOutdoorSeating <- as.factor(house$HasOutdoorSeating)
house$HasLotFrontage <- as.factor(house$HasLotFrontage)
house$HasMasVnr <- as.factor(house$HasMasVnr)
house <- na.omit(house)

# Visualizing bivariate plots, univariate plots, and correlation
ggpairs(house, c(3, 28, 29, 45))

# Looking for transformations
summary(a1<-powerTransform(cbind(LotArea,X1stFlrSF,GrLivArea) ~ 1, house))

# Double checking with marginal tests
with(house,invTranPlot(LotArea,log(SalePrice),lambda=c(-1,0,1)))
with(house,invTranPlot(X1stFlrSF,log(SalePrice),lambda=c(-1,0,1)))
with(house,invTranPlot(GrLivArea,log(SalePrice),lambda=c(-1,0,1)))

# Double checking transformation choice for response
m1 <- lm(SalePrice~log(LotArea)+log(X1stFlrSF)+log(GrLivArea),data=house)
# First, we look at an inverse response plot
inverseResponsePlot(m1)
# We see that lambda = 0.13 best fits the data

# Now we look at the Box-Cox plot
boxCox(m1)
```

```
# This summary table from the power test confirms our thoughts from
# the previous two graphs
summary(powerTransform(m1))

# All log transform for LotArea, X1stFlrSF, GrLivArea
pairs(~log(LotArea)+log(X1stFlrSF)+log(GrLivArea)+log(SalePrice),house)
testTransform(a1,c(0,0,0))

# Testing correlations of transformed predictors with transformed response
with(house, cor.test(log(LotArea),log(SalePrice)))
with(house, cor.test(log(X1stFlrSF),log(SalePrice)))
with(house, cor.test(log(GrLivArea),log(SalePrice)))

# Visualize transformed data
house$logLotArea <- log(house$LotArea)
house$logX1stFlrSF <- log(house$X1stFlrSF)
house$logGrLivArea <- log(house$GrLivArea)
house$logSalePrice <- log(house$SalePrice)
ggpairs(house, c(53, 54, 55, 56))

# Variable selection ----
# There are 52 variables total and now we wish to do variable selection
# 40 were factors and 12 numeric of some sort
# Now performing model selection
m0 <- lm(log(SalePrice) ~ 1, house) # the base model
f <- ~MSSubClass+MSZoning+log(LotArea)+LotShape+LandContour+LotConfig+LandSlope+Neighborhood+Condition1+

# Forward selection with BIC
m.forward <- step(m0, scope = f, direction = "forward", k = log(dim(house)[1]))
# Gives this model
# log(SalePrice) ~ OverallQual + log(GrLivArea) + Neighborhood +
#  MSSubClass + OverallCond + log(LotArea) + HasBasement + RoofMatl +
#  YearRemodAdd + HasGarage + Fireplaces + MSZoning + KitchenAbvGr +
#  Functional + HasX2ndFlr + SaleCondition + KitchenQual + Condition2 +
#  CentralAir + BedroomAbvGr

# Model diagnostics ----
m1 <- lm(log(SalePrice) ~ OverallQual + log(GrLivArea) + Neighborhood +
           MSSubClass + OverallCond + log(LotArea) + HasBasement + RoofMatl +
           YearRemodAdd + HasGarage + Fireplaces + MSZoning + KitchenAbvGr +
           Functional + HasX2ndFlr + SaleCondition + KitchenQual + Condition2 +
           CentralAir + BedroomAbvGr, house)
Anova(m1)
summary(m1)

# Testing for Outliers
outlierTest(m1)
# There are 10 outliers at the 0.05 Bonferroni corrected significance level
plot(m1,which = c(4))
# Examine cooks distance
cdm1 <- cooks.distance(m1)
cdm1[(cdm1) >= .35 | is.na(cdm1)]
```

```r
# Model diagnostics with influential points removed ----
# Let's start with removing 524 and 826
m2 <- lm(log(SalePrice) ~ OverallQual + log(GrLivArea) + Neighborhood +
           MSSubClass + OverallCond + log(LotArea) + HasBasement + RoofMatl +
           YearRemodAdd + HasGarage + Fireplaces + MSZoning + KitchenAbvGr +
           Functional + HasX2ndFlr + SaleCondition + KitchenQual + Condition2 +
           CentralAir + BedroomAbvGr, house, subset = -c(524, 826, 121, 272, 376, 524, 534, 584,
                                                          667, 826, 1004, 1231, 1276, 1299))
summary(m2)

# Still some outliers
outlierTest(m2)
# Nothing is particularily influential
plot(m2,which = c(1,2,3,4))

# Examine cooks distance again
cdm2 <- cooks.distance(m2)
cdm2[(cdm2) >= 0.1 | is.na(cdm2)]
# The result is nothing

# Looking at Anova for m2
Anova(m2)
# Wow, Condition2 and RoofMatl is no longer significant!

# Creating model to see if can remove the above two terms
m3 <- lm(log(SalePrice) ~ OverallQual + log(GrLivArea) + Neighborhood +
           MSSubClass + OverallCond + log(LotArea) + HasBasement +
           YearRemodAdd + HasGarage + Fireplaces + MSZoning + KitchenAbvGr +
           Functional + HasX2ndFlr + SaleCondition + KitchenQual + Condition2+
           CentralAir + BedroomAbvGr, house, subset = -c(524, 826, 121, 272, 376, 524, 534, 584,
                                                          667, 826, 1004, 1231, 1276, 1299))
# Performing F test
anova(m3,m2)
# Good to remove RoofMatl
Anova(m3)
# Looks like Condition2 is still not significant in the model
m4 <- lm(log(SalePrice) ~ OverallQual + log(GrLivArea) + Neighborhood +
           MSSubClass + OverallCond + log(LotArea) + HasBasement +
           YearRemodAdd + HasGarage + Fireplaces + MSZoning + KitchenAbvGr +
           Functional + HasX2ndFlr + SaleCondition + KitchenQual +
           CentralAir + BedroomAbvGr, house, subset = -c(524, 826, 121, 272, 376, 524, 534, 584,
                                                          667, 826, 1004, 1231, 1276, 1299))


# Performing F test
anova(m4,m3)
# We are good to remove Condition2 from model as well

# Examining this reduced model
Anova(m4)
# Looking at residual plot for possibly misspecified mean function
residualPlots(m4)
# There is curvature in the log(GrLivArea) residual plot, while everything else
```

```r
# looks fine. We acknowledge this shortcoming in our model, perhaps this suggests
# that the model fit is not the best

# Now testing for outliers
outlierTest(m4)
# Still a few outliers

# Looking for influential points
plot(m4,c(4))
# No single point appears to have a Cook's distance that is particularily larger
# than all the other points, so we have no additional points to remove

# Testing for nonconstant variance
ncvTest(m4)
# Very strong evidence against constant variance.
# We decided to use a sandwich estimator to accomdate this misspecified variance
# due to the large number of variables in our model that could possibly be
# contributing to nonconstant variance

# Checking normality assumption
plot(m4,c(2))
# Residuals do not appear to be normally distributed, since there is some
# strong curvature for the left end of the normality plot. However, our sample
# size is very large, and in general regression is robust toward violations of
# normality, so we simply acknowledge this pitfall in this model.

# Here is the regression output
summary(m4)


# Thus, we display the results of our final model
Anova(m4,vcov. = hccm)
coeftest(m4, vcov. =hccm)
coefci(m4,vcov. = hccm)

# Making effects plot for interpretation
efffects <- allEffects(m4, vcov. = hccm)
plot(efffects, ask=TRUE, multiline=TRUE, rug=FALSE, grid=TRUE,
     ci.style="bars", key.arg=list(corner=c(.975, .025)))
```

Appendix B

Data Dictionary

MSSubClass: Identifies the type of dwelling involved in the sale.

```
20  1-STORY 1946 & NEWER ALL STYLES
30  1-STORY 1945 & OLDER
40  1-STORY W/FINISHED ATTIC ALL AGES
45  1-1/2 STORY - UNFINISHED ALL AGES
50  1-1/2 STORY FINISHED ALL AGES
60  2-STORY 1946 & NEWER
```

```
 70  2-STORY 1945 & OLDER
 75  2-1/2 STORY ALL AGES
 80  SPLIT OR MULTI-LEVEL
 85  SPLIT FOYER
 90  DUPLEX - ALL STYLES AND AGES
120  1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150  1-1/2 STORY PUD - ALL AGES
160  2-STORY PUD - 1946 & NEWER
180  PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190  2 FAMILY CONVERSION - ALL STYLES AND AGES
```

MSZoning: Identifies the general zoning classification of the sale.

```
A    Agriculture
C    Commercial
FV   Floating Village Residential
I    Industrial
RH   Residential High Density
RL   Residential Low Density
RP   Residential Low Density Park
RM   Residential Medium Density
```

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

```
Grvl Gravel
Pave Paved
```

Alley: Type of alley access to property

```
Grvl Gravel
Pave Paved
NA   No alley access
```

LotShape: General shape of property

```
Reg  Regular
IR1  Slightly irregular
IR2  Moderately Irregular
IR3  Irregular
```

LandContour: Flatness of the property

```
Lvl  Near Flat/Level
Bnk  Banked - Quick and significant rise from street grade to building
HLS  Hillside - Significant slope from side to side
Low  Depression
```

Utilities: Type of utilities available

```
AllPub   All public Utilities (E,G,W,& S)
NoSewr   Electricity, Gas, and Water (Septic Tank)
NoSeWa   Electricity and Gas Only
ELO  Electricity only
```

LotConfig: Lot configuration

```
Inside   Inside lot
Corner   Corner lot
CulDSac  Cul-de-sac
FR2  Frontage on 2 sides of property
FR3  Frontage on 3 sides of property
```

LandSlope: Slope of property

```
Gtl  Gentle slope
Mod  Moderate Slope
Sev  Severe Slope
```

Neighborhood: Physical locations within Ames city limits

```
Blmngtn  Bloomington Heights
Blueste  Bluestem
BrDale   Briardale
BrkSide  Brookside
ClearCr  Clear Creek
CollgCr  College Creek
Crawfor  Crawford
Edwards  Edwards
Gilbert  Gilbert
IDOTRR   Iowa DOT and Rail Road
MeadowV  Meadow Village
Mitchel  Mitchell
Names    North Ames
NoRidge  Northridge
NPkVill  Northpark Villa
NridgHt  Northridge Heights
NWAmes   Northwest Ames
OldTown  Old Town
SWISU    South & West of Iowa State University
Sawyer   Sawyer
SawyerW  Sawyer West
Somerst  Somerset
StoneBr  Stone Brook
Timber   Timberland
Veenker  Veenker
```

Condition1: Proximity to various conditions

```
Artery   Adjacent to arterial street
Feedr    Adjacent to feeder street
Norm Normal
RRNn Within 200' of North-South Railroad
RRAn Adjacent to North-South Railroad
PosN Near positive off-site feature--park, greenbelt, etc.
PosA Adjacent to postive off-site feature
RRNe Within 200' of East-West Railroad
RRAe Adjacent to East-West Railroad
```

Condition2: Proximity to various conditions (if more than one is present)

```
Artery   Adjacent to arterial street
Feedr    Adjacent to feeder street
Norm Normal
RRNn Within 200' of North-South Railroad
RRAn Adjacent to North-South Railroad
PosN Near positive off-site feature--park, greenbelt, etc.
PosA Adjacent to postive off-site feature
RRNe Within 200' of East-West Railroad
RRAe Adjacent to East-West Railroad
```

BldgType: Type of dwelling

```
1Fam Single-family Detached
2FmCon   Two-family Conversion; originally built as one-family dwelling
Duplx    Duplex
TwnhsE   Townhouse End Unit
TwnhsI   Townhouse Inside Unit
```

HouseStyle: Style of dwelling

```
1Story   One story
1.5Fin   One and one-half story: 2nd level finished
1.5Unf   One and one-half story: 2nd level unfinished
2Story   Two story
2.5Fin   Two and one-half story: 2nd level finished
2.5Unf   Two and one-half story: 2nd level unfinished
SFoyer   Split Foyer
SLvl Split Level
```

OverallQual: Rates the overall material and finish of the house

```
10   Very Excellent
9    Excellent
8    Very Good
7    Good
6    Above Average
5    Average
4    Below Average
3    Fair
2    Poor
1    Very Poor
```

OverallCond: Rates the overall condition of the house

```
10   Very Excellent
9    Excellent
8    Very Good
7    Good
6    Above Average
5    Average
4    Below Average
3    Fair
2    Poor
1    Very Poor
```

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

```
Flat Flat
Gable    Gable
Gambrel  Gabrel (Barn)
Hip  Hip
Mansard  Mansard
Shed Shed
```

RoofMatl: Roof material

```
ClyTile  Clay or Tile
CompShg  Standard (Composite) Shingle
Membran  Membrane
Metal    Metal
Roll Roll
Tar&Grv  Gravel & Tar
WdShake  Wood Shakes
WdShngl  Wood Shingles
```

Exterior1st: Exterior covering on house

```
AsbShng  Asbestos Shingles
AsphShn  Asphalt Shingles
BrkComm  Brick Common
BrkFace  Brick Face
CBlock   Cinder Block
CemntBd  Cement Board
HdBoard  Hard Board
ImStucc  Imitation Stucco
MetalSd  Metal Siding
Other    Other
Plywood  Plywood
```

```
PreCast  PreCast
Stone    Stone
Stucco   Stucco
VinylSd  Vinyl Siding
Wd Sdng  Wood Siding
WdShing  Wood Shingles
```

Exterior2nd: Exterior covering on house (if more than one material)

```
AsbShng  Asbestos Shingles
AsphShn  Asphalt Shingles
BrkComm  Brick Common
BrkFace  Brick Face
CBlock   Cinder Block
CemntBd  Cement Board
HdBoard  Hard Board
ImStucc  Imitation Stucco
MetalSd  Metal Siding
Other    Other
Plywood  Plywood
PreCast  PreCast
Stone    Stone
Stucco   Stucco
VinylSd  Vinyl Siding
Wd Sdng  Wood Siding
WdShing  Wood Shingles
```

MasVnrType: Masonry veneer type

```
BrkCmn   Brick Common
BrkFace  Brick Face
CBlock   Cinder Block
None None
Stone    Stone
```

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

```
Ex   Excellent
Gd   Good
TA   Average/Typical
Fa   Fair
Po   Poor
```

ExterCond: Evaluates the present condition of the material on the exterior

```
Ex   Excellent
Gd   Good
TA   Average/Typical
Fa   Fair
Po   Poor
```

Foundation: Type of foundation

```
BrkTil   Brick & Tile
CBlock   Cinder Block
PConc    Poured Contrete
Slab Slab
Stone    Stone
Wood Wood
```

BsmtQual: Evaluates the height of the basement

```
Ex   Excellent (100+ inches)
Gd   Good (90-99 inches)
TA   Typical (80-89 inches)
Fa   Fair (70-79 inches)
Po   Poor (<70 inches
NA   No Basement
```

BsmtCond: Evaluates the general condition of the basement

```
Ex   Excellent
Gd   Good
TA   Typical - slight dampness allowed
Fa   Fair - dampness or some cracking or settling
Po   Poor - Severe cracking, settling, or wetness
NA   No Basement
```

BsmtExposure: Refers to walkout or garden level walls

```
Gd   Good Exposure
Av   Average Exposure (split levels or foyers typically score average or above)
Mn   Mimimum Exposure
No   No Exposure
NA   No Basement
```

BsmtFinType1: Rating of basement finished area

```
GLQ  Good Living Quarters
ALQ  Average Living Quarters
BLQ  Below Average Living Quarters
Rec  Average Rec Room
LwQ  Low Quality
Unf  Unfinshed
NA   No Basement
```

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

```
GLQ  Good Living Quarters
ALQ  Average Living Quarters
BLQ  Below Average Living Quarters
Rec  Average Rec Room
LwQ  Low Quality
Unf  Unfinshed
NA   No Basement
```

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

```
Floor    Floor Furnace
GasA Gas forced warm air furnace
GasW Gas hot water or steam heat
Grav Gravity furnace
OthW Hot water or steam heat other than gas
Wall Wall furnace
```

HeatingQC: Heating quality and condition

```
Ex   Excellent
Gd   Good
TA   Average/Typical
Fa   Fair
Po   Poor
```

CentralAir: Central air conditioning

```
N    No
Y    Yes
```

Electrical: Electrical system

```
SBrkr    Standard Circuit Breakers & Romex
FuseA    Fuse Box over 60 AMP and all Romex wiring (Average)
FuseF    60 AMP Fuse Box and mostly Romex wiring (Fair)
FuseP    60 AMP Fuse Box and mostly knob & tube wiring (poor)
Mix  Mixed
```

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

```
Ex   Excellent
Gd   Good
TA   Typical/Average
Fa   Fair
Po   Poor
```

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)

```
Typ  Typical Functionality
Min1 Minor Deductions 1
Min2 Minor Deductions 2
Mod  Moderate Deductions
Maj1 Major Deductions 1
Maj2 Major Deductions 2
Sev  Severely Damaged
Sal  Salvage only
```

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

```
Ex   Excellent - Exceptional Masonry Fireplace
Gd   Good - Masonry Fireplace in main level
TA   Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement
Fa   Fair - Prefabricated Fireplace in basement
Po   Poor - Ben Franklin Stove
NA   No Fireplace
```

GarageType: Garage location

```
2Types   More than one type of garage
Attchd   Attached to home
Basment  Basement Garage
BuiltIn  Built-In (Garage part of house - typically has room above garage)
CarPort  Car Port
Detchd   Detached from home
NA   No Garage
```

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

```
Fin  Finished
RFn  Rough Finished
Unf  Unfinished
NA   No Garage
```

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

```
Ex   Excellent
Gd   Good
TA   Typical/Average
Fa   Fair
Po   Poor
NA   No Garage
```

GarageCond: Garage condition

```
Ex   Excellent
Gd   Good
TA   Typical/Average
Fa   Fair
Po   Poor
NA   No Garage
```

PavedDrive: Paved driveway

```
Y    Paved
P    Partial Pavement
N    Dirt/Gravel
```

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

```
Ex   Excellent
Gd   Good
TA   Average/Typical
Fa   Fair
NA   No Pool
```

Fence: Fence quality

```
GdPrv    Good Privacy
MnPrv    Minimum Privacy
GdWo Good Wood
MnWw Minimum Wood/Wire
NA   No Fence
```

MiscFeature: Miscellaneous feature not covered in other categories

```
Elev Elevator
Gar2 2nd Garage (if not described in garage section)
Othr Other
Shed Shed (over 100 SF)
TenC Tennis Court
NA   None
```

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

```
WD   Warranty Deed - Conventional
CWD  Warranty Deed - Cash
VWD  Warranty Deed - VA Loan
New  Home just constructed and sold
COD  Court Officer Deed/Estate
Con  Contract 15% Down payment regular terms
ConLw    Contract Low Down payment and low interest
ConLI    Contract Low Interest
ConLD    Contract Low Down
Oth  Other
```

SaleCondition: Condition of sale

```
Normal   Normal Sale
Abnorml  Abnormal Sale -  trade, foreclosure, short sale
AdjLand  Adjoining Land Purchase
Alloca   Allocation - two linked properties with separate deeds, typically condo with a garage unit
Family   Sale between family members
Partial  Home was not completed when last assessed (associated with New Homes)
```

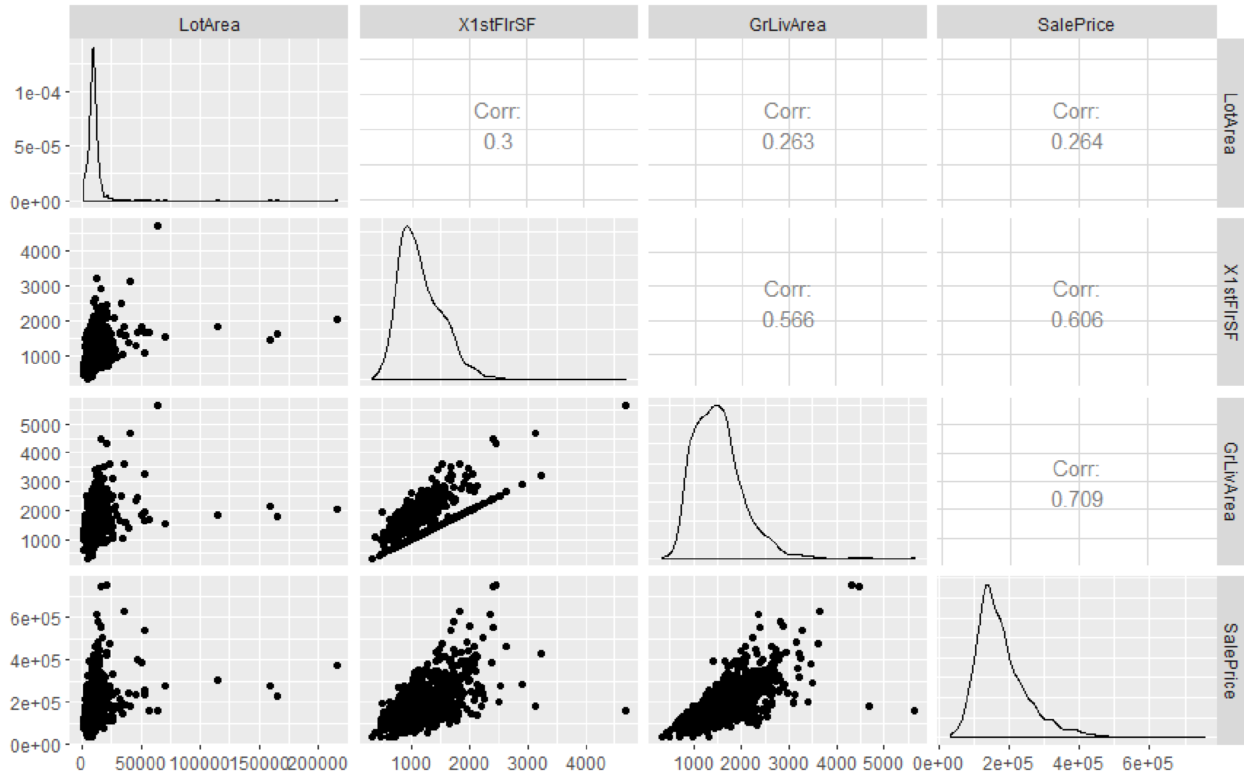


Figure 1

```
bcPower Transformations to Multinormality
          Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
LotArea      0.0209           0      -0.0234       0.0651
X1stFlrSF   -0.0392           0      -0.1609       0.0825
GrLivArea    0.0158           0      -0.0985       0.1301

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
                               LRT df    pval
LR test, lambda = (0 0 0) 1.412987  3 0.70249

Likelihood ratio test that no transformations are needed
                               LRT df       pval
LR test, lambda = (1 1 1) 2966.806  3 < 2.22e-16
```

Figure 2

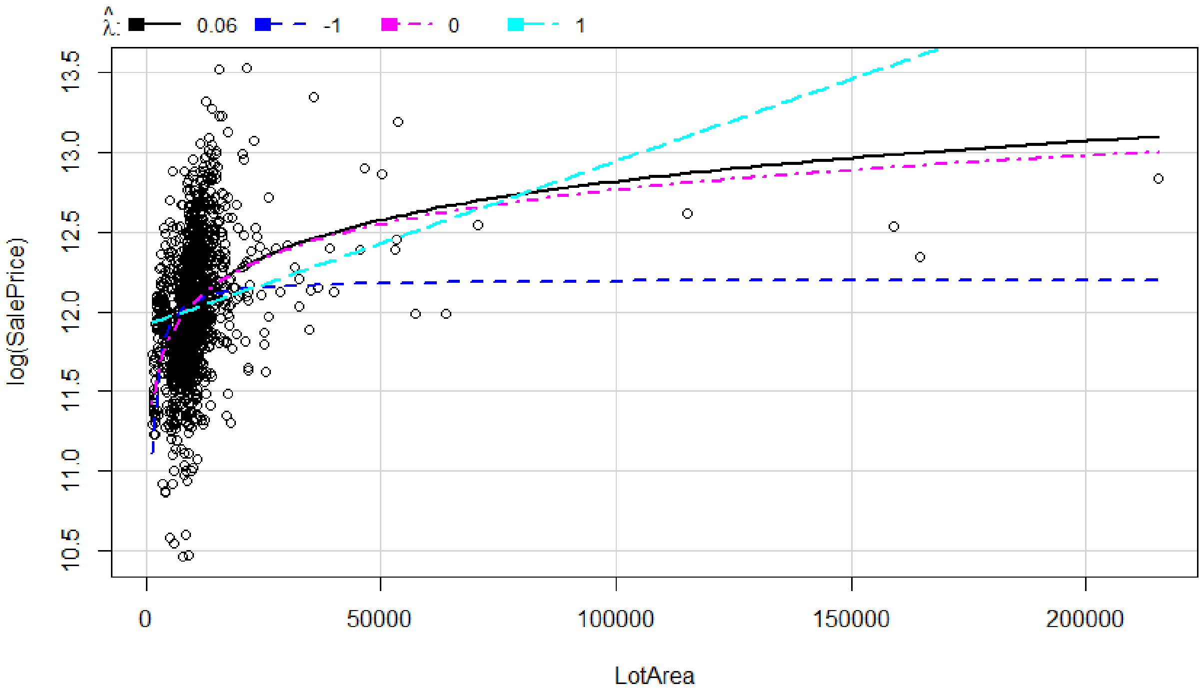


Figure 3

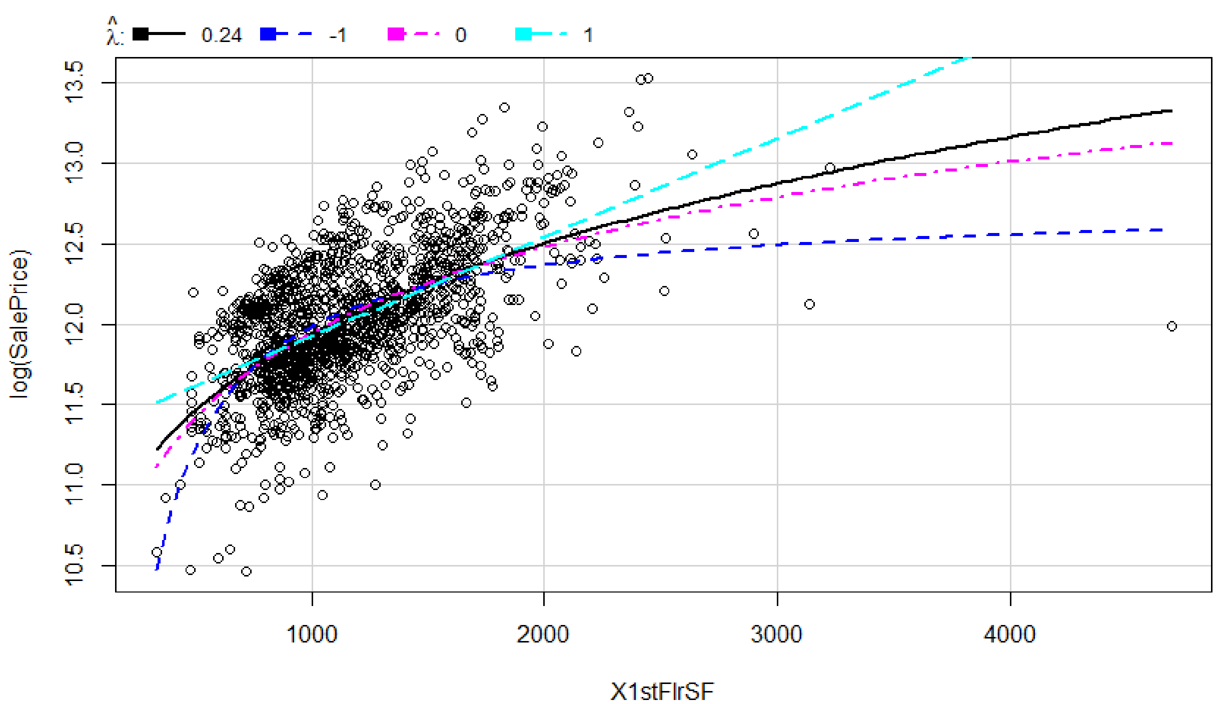


Figure 4

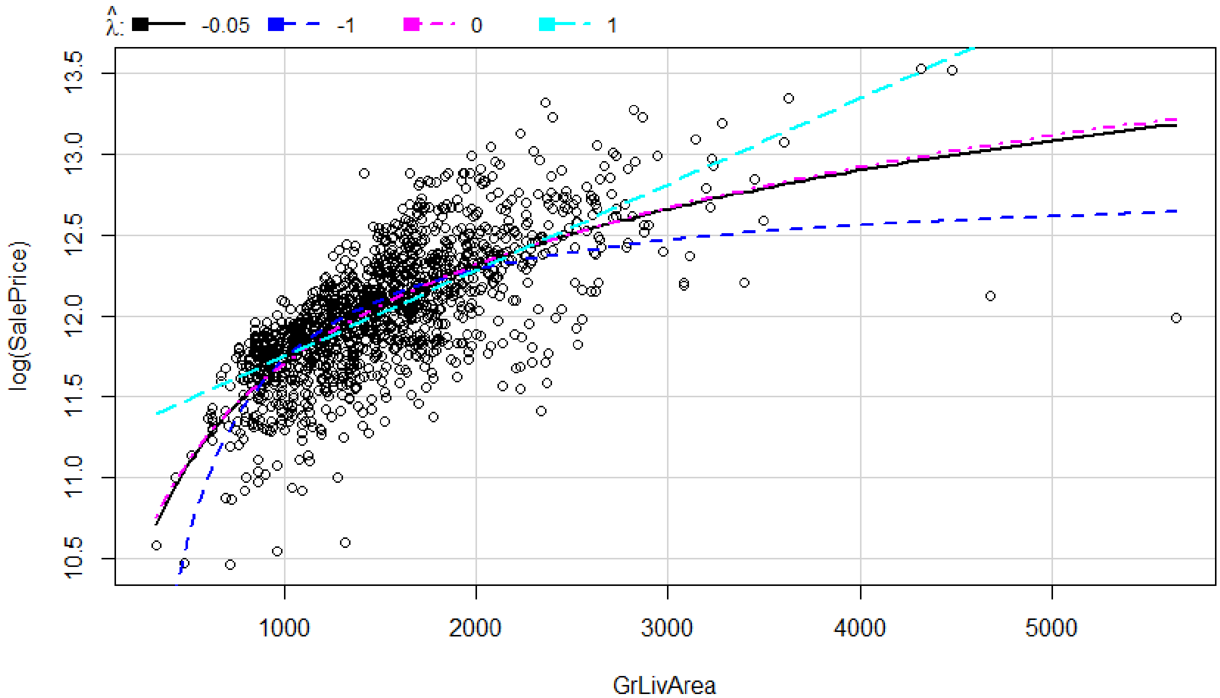

λ̂: -0.05 -1 0 1
log(SalePrice)
10.5
11.0
11.5
12.0
12.5
13.0
13.5
1000
2000
3000
4000
5000
GrLivArea

Figure 5

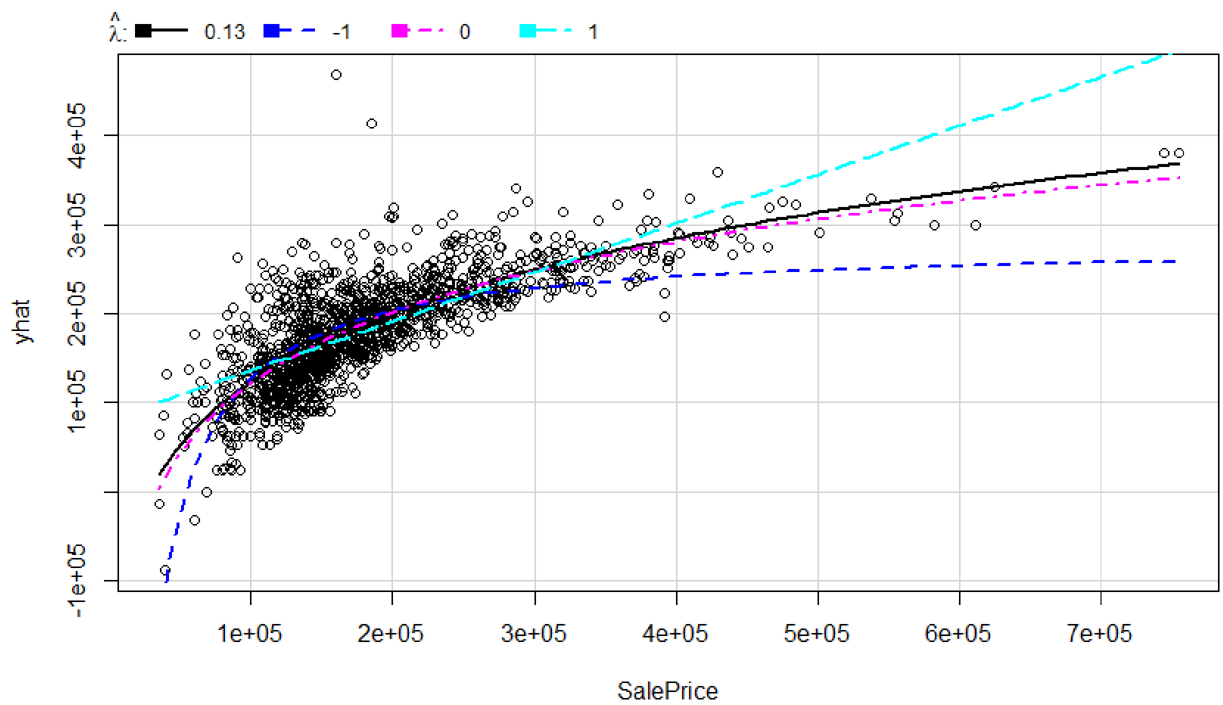


Figure 6

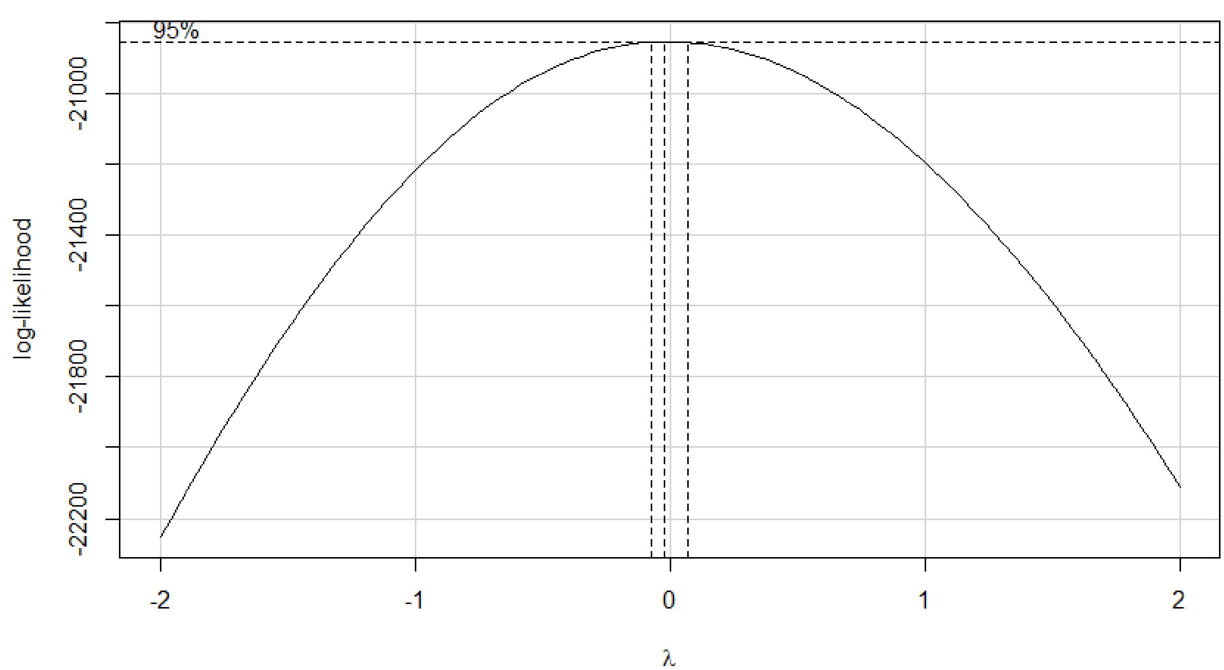


Figure 7

```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    -0.004           0      -0.0771        0.069

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                             LRT df    pval
LR test, lambda = (0) 0.01170519  1 0.91384

Likelihood ratio test that no transformation is needed
                           LRT df       pval
LR test, lambda = (1) 683.0536  1 < 2.22e-16
```

Figure 8

```
        Pearson's product-moment correlation

data:  log(LotArea) and log(SalePrice)
t = 16.655, df = 1457, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3559040 0.4421643
sample estimates:
      cor
0.3999193

> with(house, cor.test(log(X1stFlrSF),log(SalePrice)))

        Pearson's product-moment correlation

data:  log(X1stFlrSF) and log(SalePrice)
t = 29.327, df = 1457, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5759469 0.6405504
sample estimates:
      cor
0.6092586

> with(house, cor.test(log(GrLivArea),log(SalePrice)))

        Pearson's product-moment correlation

data:  log(GrLivArea) and log(SalePrice)
t = 40.801, df = 1457, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7053725 0.7533443
sample estimates:
      cor
0.7302573
```

Figure 9

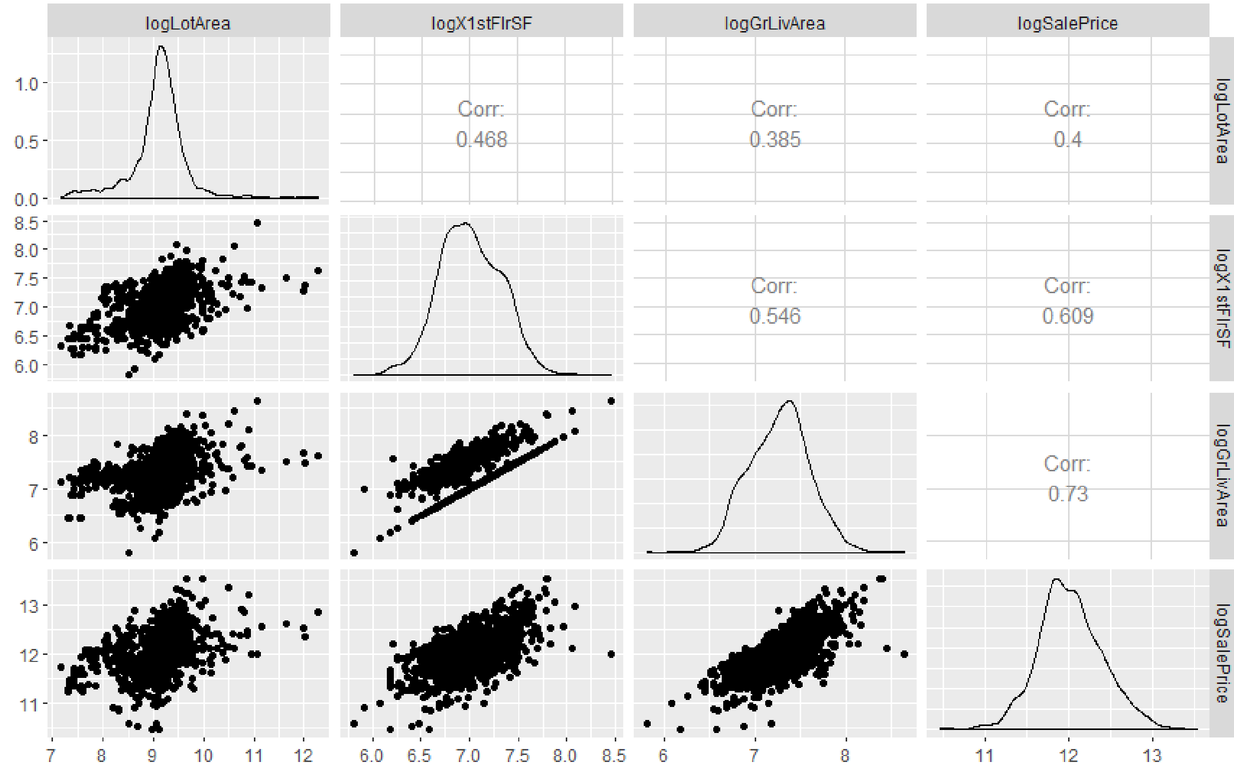


Figure 10

```
lm(formula = log(SalePrice) ~ OverallQual + log(GrLivArea) +
    Neighborhood + MSSubClass + OverallCond + log(LotArea) +
    HasBasement + RoofMatl + YearRemodAdd + HasGarage + Fireplaces +
    MSZoning + KitchenAbvGr + Functional + HasX2ndFlr + SaleCondition +
    KitchenQual + Condition2 + CentralAir + BedroomAbvGr, data = house)
```

Figure 11

	rstudent	unadjusted p-value	Bonferroni p
826	6.110154	1.2974e-09	1.8799e-06
524	-6.110154	1.2974e-09	1.8799e-06
633	-6.031533	2.0897e-09	3.0280e-06
1325	-5.591016	2.7247e-08	3.9481e-05
969	-5.448577	6.0204e-08	8.7235e-05
589	-5.444317	6.1630e-08	8.9303e-05
463	-5.024902	5.7065e-07	8.2688e-04
31	-4.901415	1.0658e-06	1.5444e-03
411	-4.700359	2.8607e-06	4.1451e-03
1433	-4.319625	1.6764e-05	2.4292e-02

Figure 12

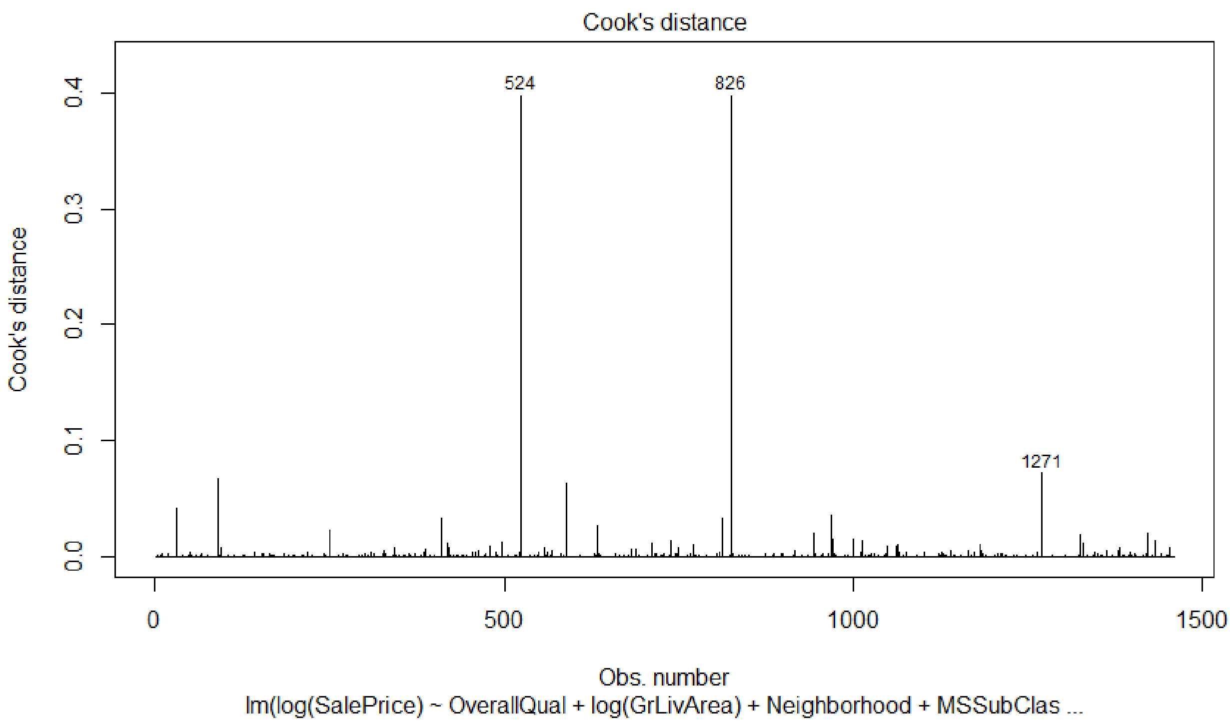


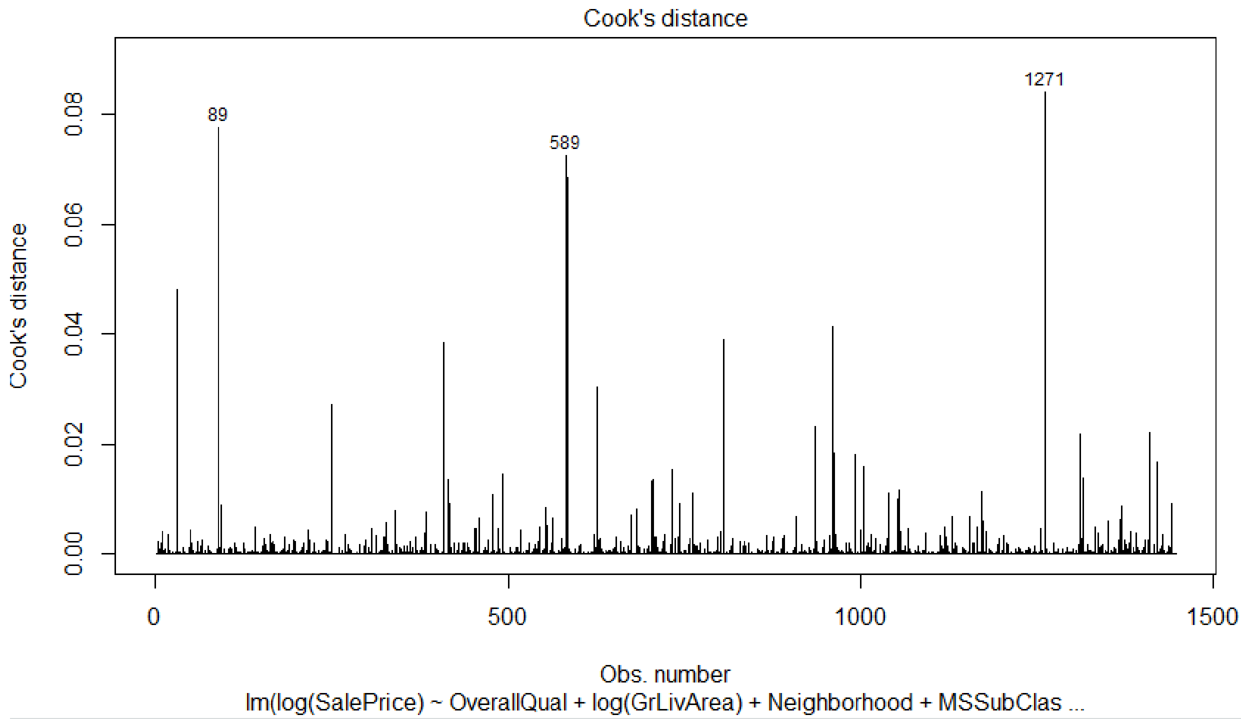
Figure 13

121	272	376	524	534	584	667	826
NaN	NaN	NaN	0.3970551	NaN	NaN	NaN	0.3970551
1004	1231	1276	1299				
NaN	NaN	NaN	NaN				

Figure 14

	rstudent	unadjusted p-value	Bonferroni p
633	-6.095696	1.4171e-09	2.0506e-06
1325	-5.679149	1.6533e-08	2.3924e-05
589	-5.509898	4.2898e-08	6.2073e-05
969	-5.488330	4.8350e-08	6.9962e-05
463	-5.067433	4.5874e-07	6.6380e-04
31	-4.958645	7.9935e-07	1.1567e-03
411	-4.795104	1.8049e-06	2.6116e-03
1433	-4.397860	1.1785e-05	1.7053e-02
971	4.358769	1.4065e-05	2.0351e-02

Figure 15



Cook's distance
Cook's distance
0.00
0.02
0.04
0.06
0.08
89
589
1271
0
500
1000
1500
Obs. number
lm(log(SalePrice) ~ OverallQual + log(GrLivArea) + Neighborhood + MSSubClas ...

Figure 16

```
Anova Table (Type II tests)

Response: log(SalePrice)
                 Sum Sq   Df  F value    Pr(>F)
OverallQual      3.0392    8  26.8010 < 2.2e-16 ***
log(GrLivArea)   9.5075    1 670.7366 < 2.2e-16 ***
Neighborhood     2.4102   24   7.0849 < 2.2e-16 ***
MSSubClass       1.1360   14   5.7247 6.057e-11 ***
OverallCond      1.1865    7  11.9576 7.214e-15 ***
log(LotArea)     0.8999    1  63.4856 3.390e-15 ***
HasBasement      0.6270    1  44.2354 4.207e-11 ***
RoofMatl         0.0489    3   1.1509 0.3273412
YearRemodAdd     0.4345    1  30.6534 3.697e-08 ***
HasGarage        0.5124    1  36.1473 2.347e-09 ***
Fireplaces       0.2636    1  18.5958 1.732e-05 ***
MSZoning         0.8298    4  14.6345 1.042e-11 ***
KitchenAbvGr     0.2719    1  19.1804 1.280e-05 ***
Functional       0.8661    5  12.2201 1.316e-11 ***
HasX2ndFlr       0.3151    1  22.2299 2.667e-06 ***
SaleCondition    0.8071    5  11.3875 8.770e-11 ***
KitchenQual      0.3852    3   9.0577 6.134e-06 ***
Condition2       0.0874    3   2.0543 0.1044787
CentralAir       0.2125    1  14.9928 0.0001130 ***
BedroomAbvGr     0.1722    1  12.1498 0.0005066 ***
Residuals       19.2776 1360
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 17

```
Analysis of Variance Table

Model 1: log(SalePrice) ~ OverallQual + log(GrLivArea) + Neighborhood +
    MSSubClass + OverallCond + log(LotArea) + HasBasement + YearRemodAdd +
    HasGarage + Fireplaces + MSZoning + KitchenAbvGr + Functional +
    HasX2ndFlr + SaleCondition + KitchenQual + Condition2 + CentralAir +
    BedroomAbvGr
Model 2: log(SalePrice) ~ OverallQual + log(GrLivArea) + Neighborhood +
    MSSubClass + OverallCond + log(LotArea) + HasBasement + RoofMatl +
    YearRemodAdd + HasGarage + Fireplaces + MSZoning + KitchenAbvGr +
    Functional + HasX2ndFlr + SaleCondition + KitchenQual + Condition2 +
    CentralAir + BedroomAbvGr
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1363 19.327
2   1360 19.278  3  0.048943 1.1509 0.3273
```

Figure 18

```
Anova Table (Type II tests)

Response: log(SalePrice)
                Sum Sq   Df  F value    Pr(>F)
OverallQual     3.1965    8  28.1791 < 2.2e-16 ***
log(GrLivArea)  9.6082    1 677.6131 < 2.2e-16 ***
Neighborhood    2.4060   24   7.0701 < 2.2e-16 ***
MSSubClass      1.1316   14   5.7005 6.931e-11 ***
OverallCond     1.1778    7  11.8661 9.563e-15 ***
log(LotArea)    0.9361    1  66.0198 9.916e-16 ***
HasBasement     0.6259    1  44.1414 4.403e-11 ***
YearRemodAdd    0.4316    1  30.4399 4.116e-08 ***
HasGarage       0.5097    1  35.9491 2.590e-09 ***
Fireplaces      0.2630    1  18.5445 1.779e-05 ***
MSZoning        0.8283    4  14.6034 1.103e-11 ***
KitchenAbvGr    0.2664    1  18.7859 1.570e-05 ***
Functional      0.8756    5  12.3502 9.769e-12 ***
HasX2ndFlr      0.3350    1  23.6292 1.303e-06 ***
SaleCondition   0.8203    5  11.5698 5.783e-11 ***
KitchenQual     0.3911    3   9.1946 5.048e-06 ***
Condition2      0.0857    3   2.0136 0.1101704
CentralAir      0.2173    1  15.3238 9.504e-05 ***
BedroomAbvGr    0.1716    1  12.0987 0.0005205 ***
Residuals      19.3266 1363
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 19

```
Analysis of Variance Table

Model 1: log(SalePrice) ~ OverallQual + log(GrLivArea) + Neighborhood +
    MSSubClass + OverallCond + log(LotArea) + HasBasement + YearRemodAdd +
    HasGarage + Fireplaces + MSZoning + KitchenAbvGr + Functional +
    HasX2ndFlr + SaleCondition + KitchenQual + CentralAir + BedroomAbvGr
Model 2: log(SalePrice) ~ OverallQual + log(GrLivArea) + Neighborhood +
    MSSubClass + OverallCond + log(LotArea) + HasBasement + YearRemodAdd +
    HasGarage + Fireplaces + MSZoning + KitchenAbvGr + Functional +
    HasX2ndFlr + SaleCondition + KitchenQual + Condition2 + CentralAir +
    BedroomAbvGr
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1366 19.412
2   1363 19.327  3  0.085657 2.0136 0.1102
```

Figure 20

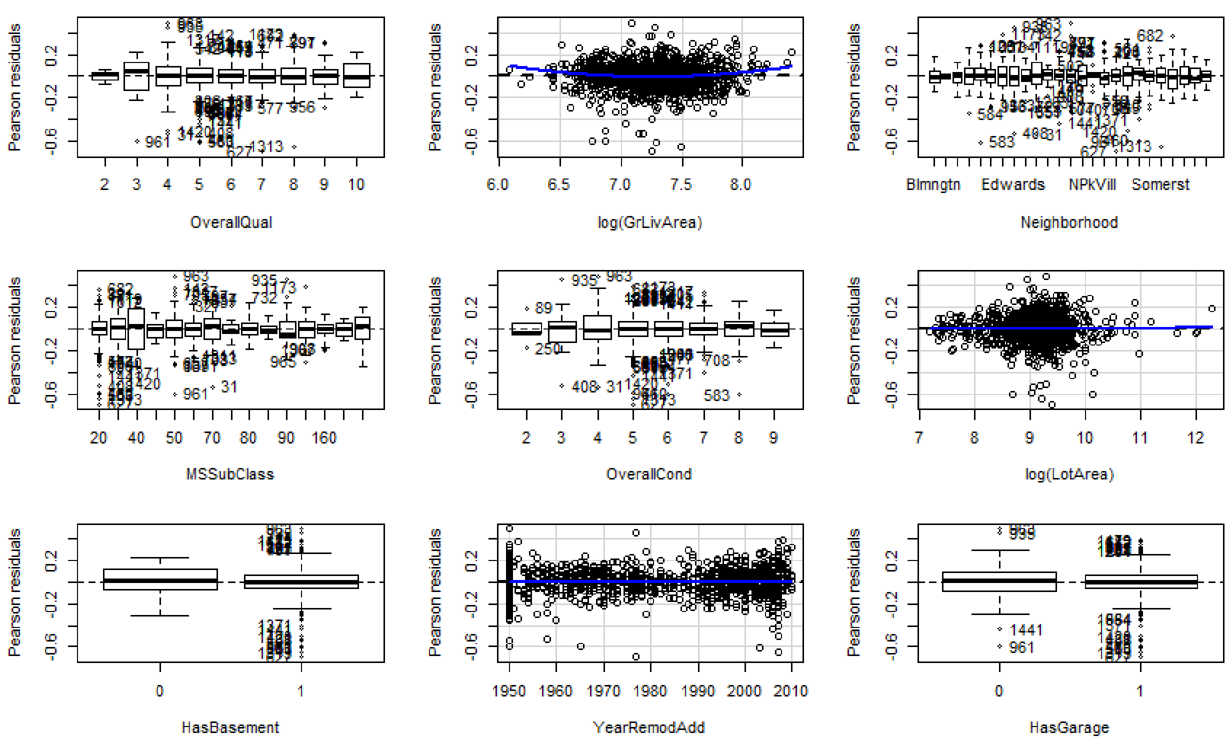


Figure 21

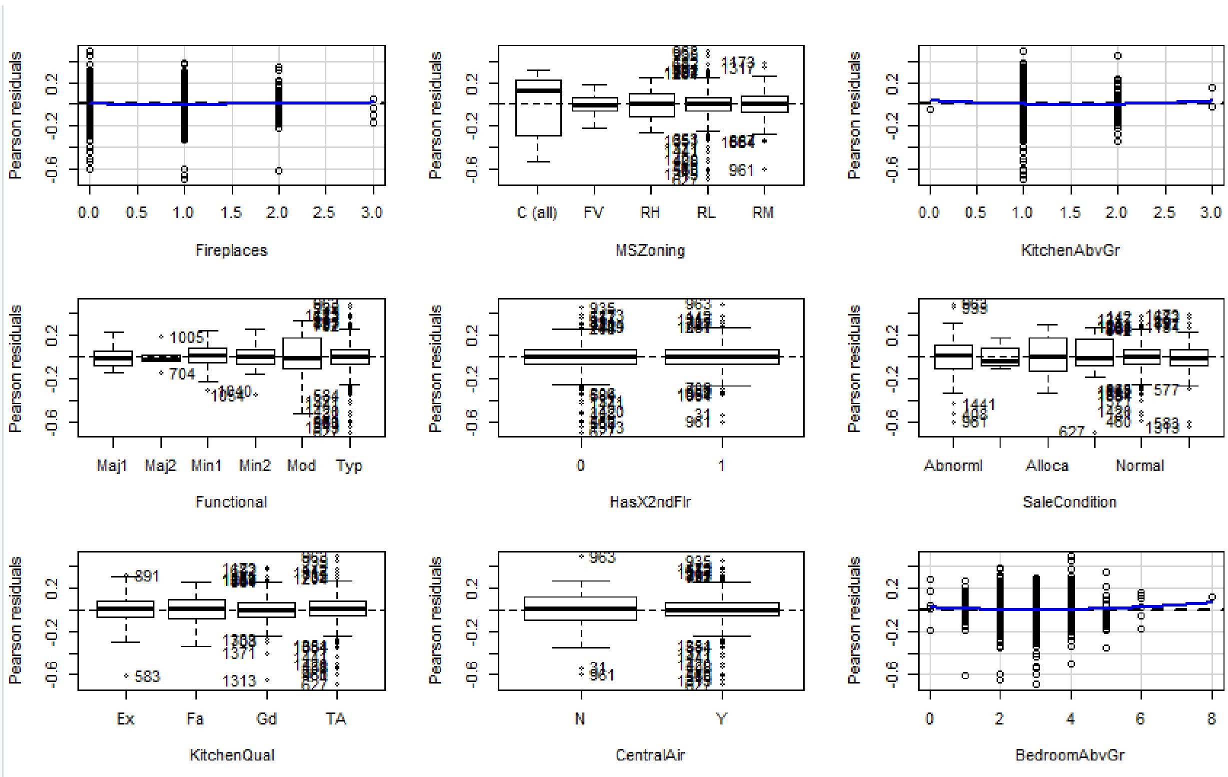


Figure 22

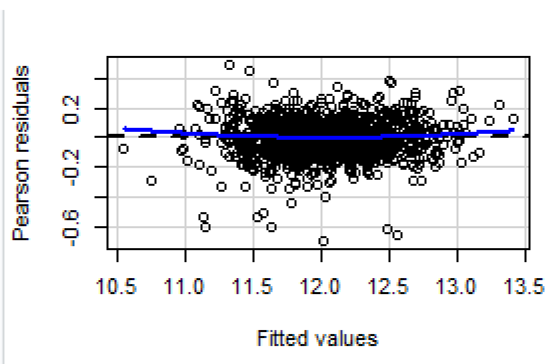


Figure 23

```
                Test stat Pr(>|Test stat|)
OverallQual
log(GrLivArea)    4.4483      9.357e-06 ***
Neighborhood
MSSubClass
OverallCond
log(LotArea)      0.4990         0.6178
HasBasement
YearRemodAdd     -0.3788         0.7049
HasGarage
Fireplaces        0.5243         0.6001
MSZoning
KitchenAbvGr      0.4735         0.6359
Functional
HasX2ndFlr
SaleCondition
KitchenQual
CentralAir
BedroomAbvGr      1.3029         0.1928
Tukey test        4.3009      1.701e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 24

```
      rstudent unadjusted p-value Bonferroni p
633  -6.093141         1.4377e-09   2.0803e-06
1325 -5.662224         1.8193e-08   2.6325e-05
589  -5.398760         7.9049e-08   1.1438e-04
969  -5.378682         8.8184e-08   1.2760e-04
463  -5.070949         4.5025e-07   6.5151e-04
31   -4.880610         1.1820e-06   1.7104e-03
411  -4.774417         1.9964e-06   2.8887e-03
1433 -4.403312         1.1493e-05   1.6630e-02
971   4.310262         1.7476e-05   2.5288e-02
```

Figure 25

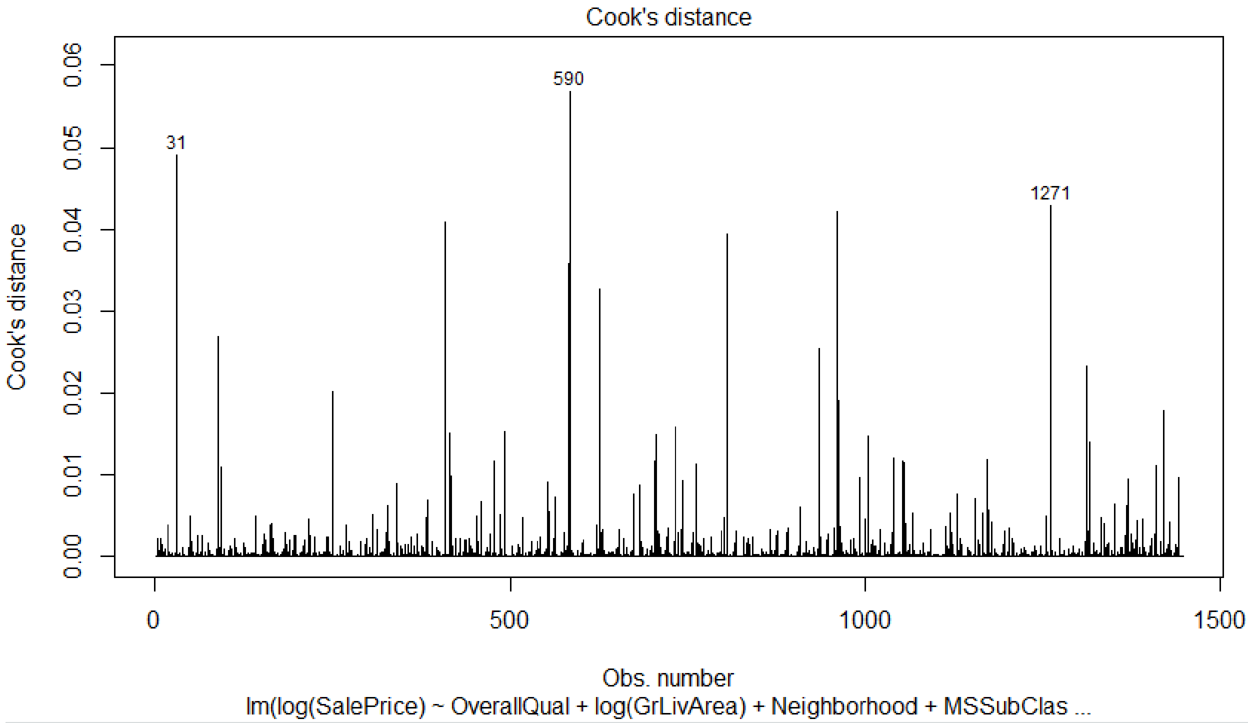

Cook's distance
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
0.06
31
590
1271
0
500
1000
1500
Obs. number
lm(log(SalePrice) ~ OverallQual + log(GrLivArea) + Neighborhood + MSSubClas ...

Figure 26

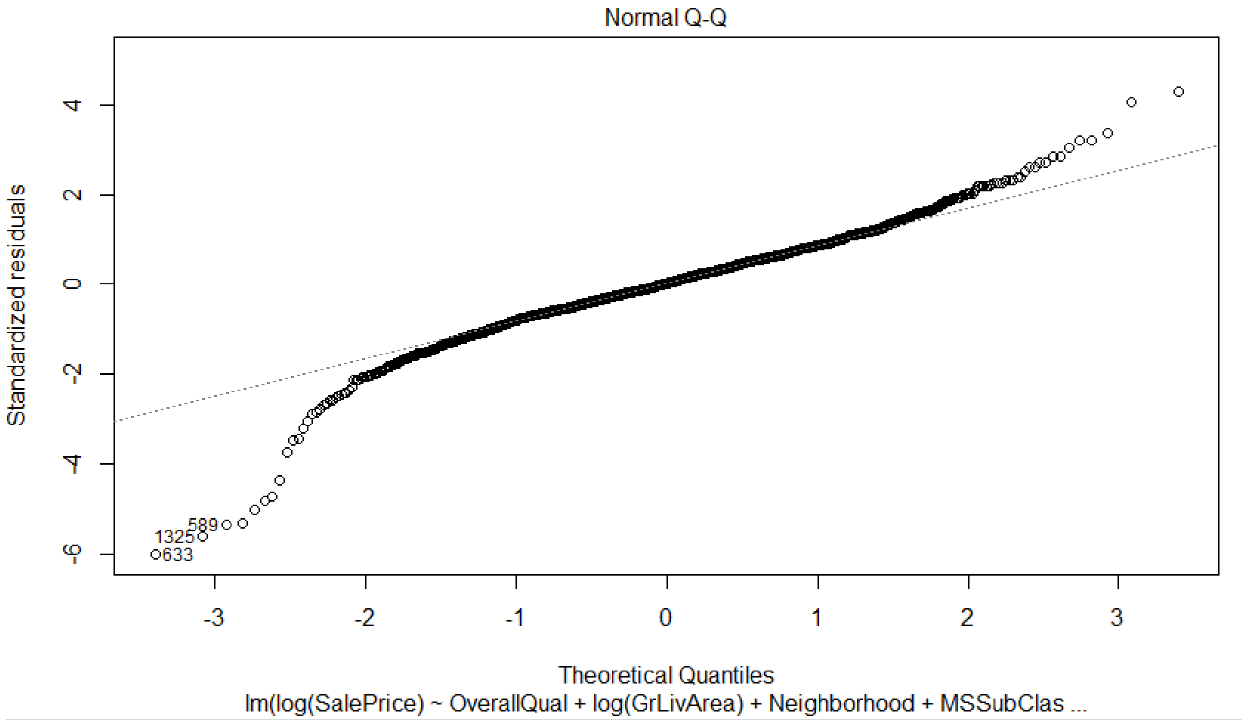


Figure 27

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 15.91483, Df = 1, p = 6.6258e-05
```

Figure 28

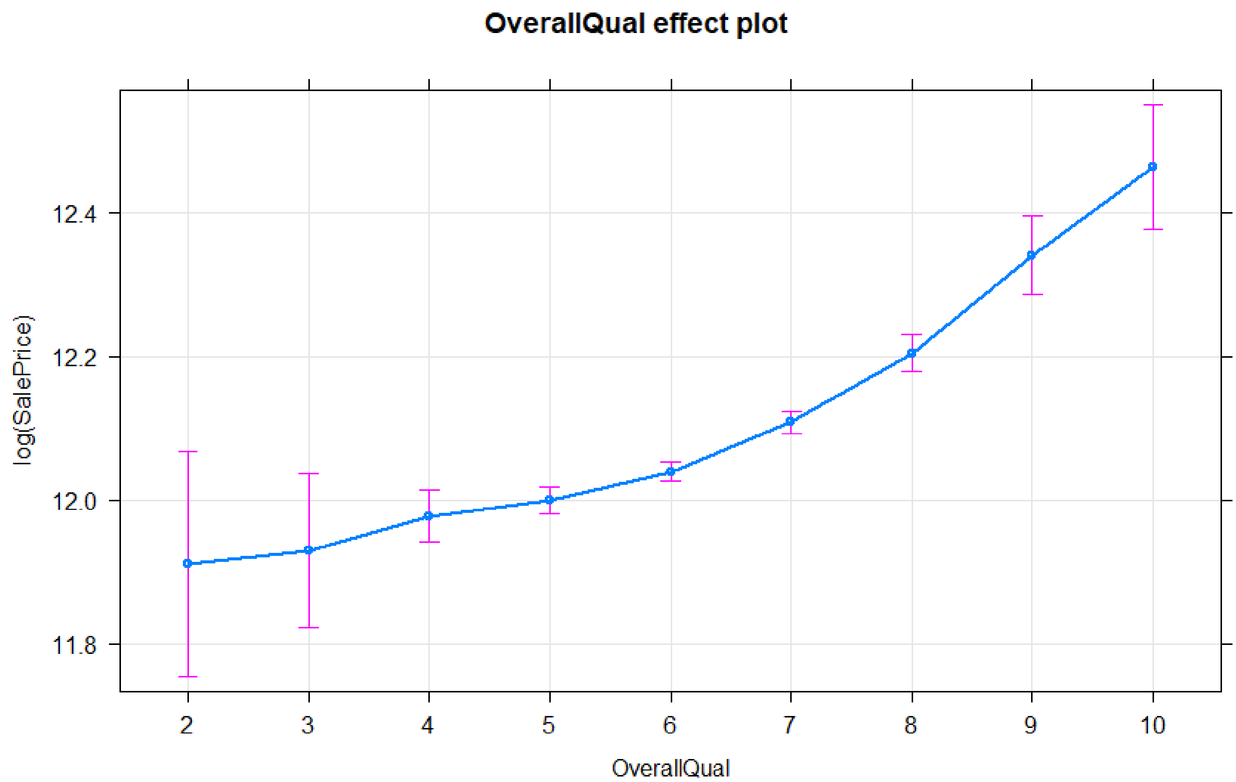

OverallQual effect plot
12.4
12.2
12.0
11.8
log(SalePrice)
2
3
4
5
6
7
8
9
10
OverallQual

Figure 29

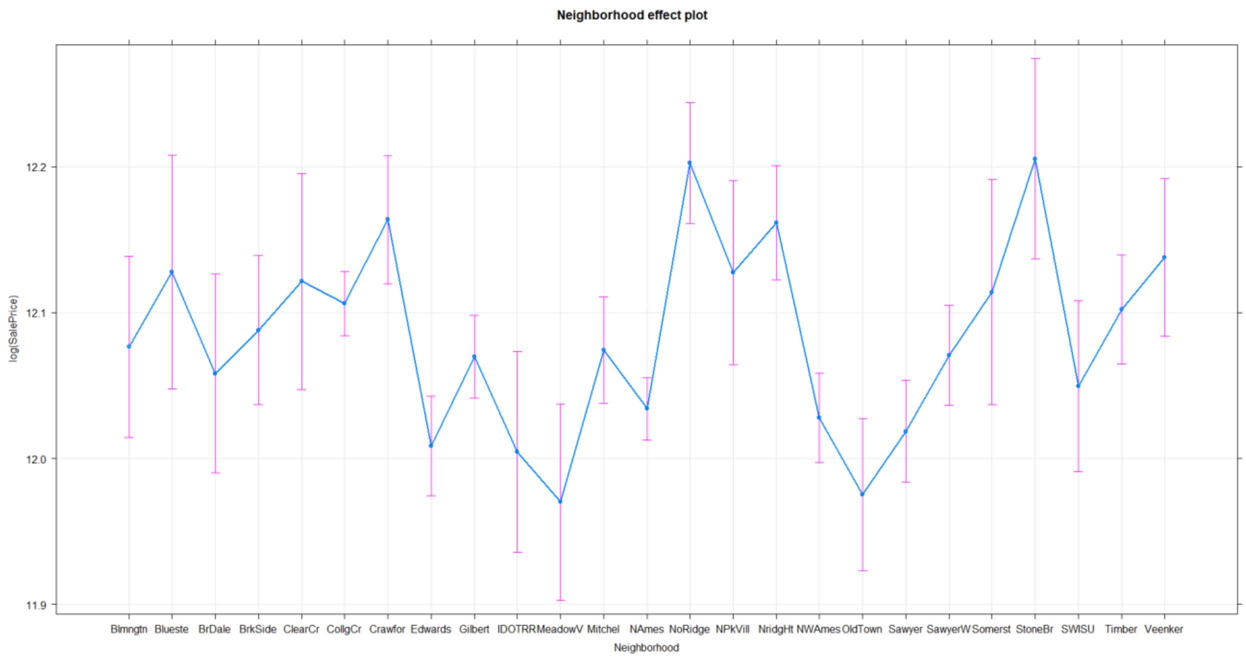

Neighborhood effect plot
log(SalePrice)
12.2
12.1
12.0
11.9
Blmngtn Blueste BrDale BrkSide ClearCr CollgCr Crawfor Edwards Gilbert IDOTRR MeadowV Mitchel NAmes NoRidge NPkVill NridgHt NWAmes OldTown Sawyer SawyerW Somerst StoneBr SWISU Timber Veenker
Neighborhood

Figure 30

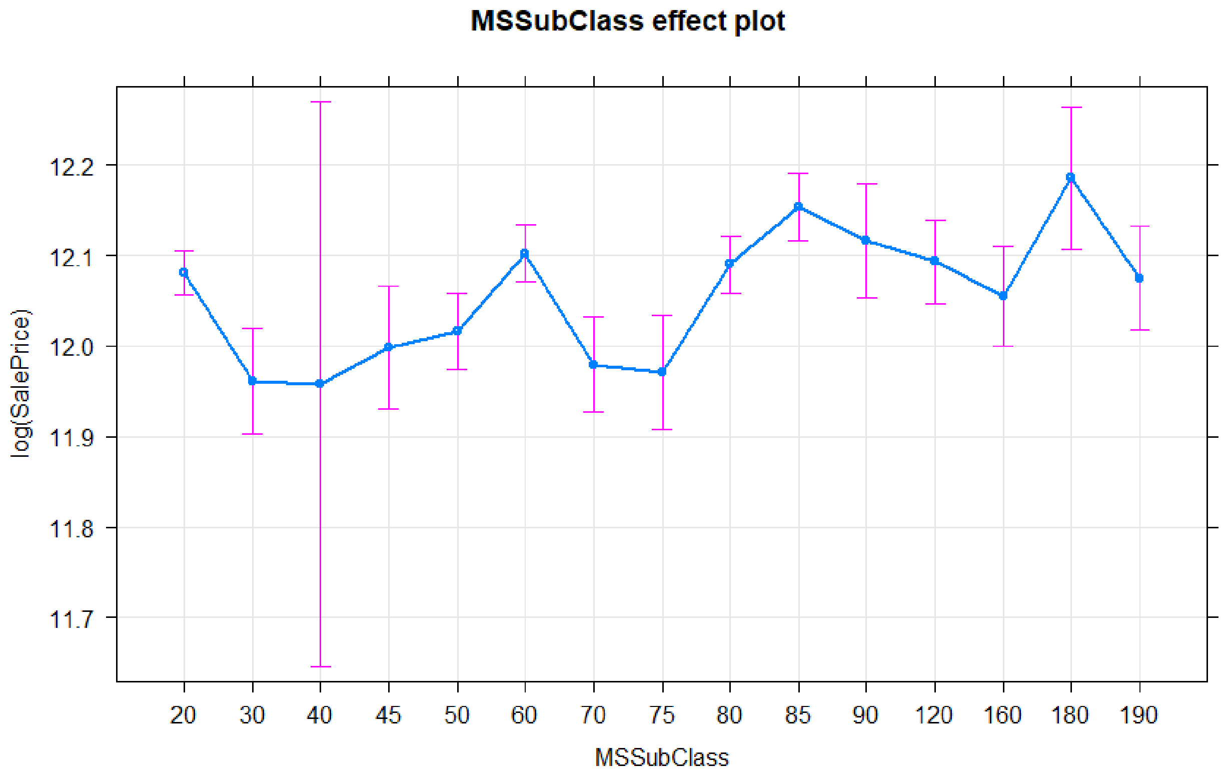

MSSubClass effect plot
log(SalePrice)
12.2
12.1
12.0
11.9
11.8
11.7
20
30
40
45
50
60
70
75
80
85
90
120
160
180
190
MSSubClass

Figure 31

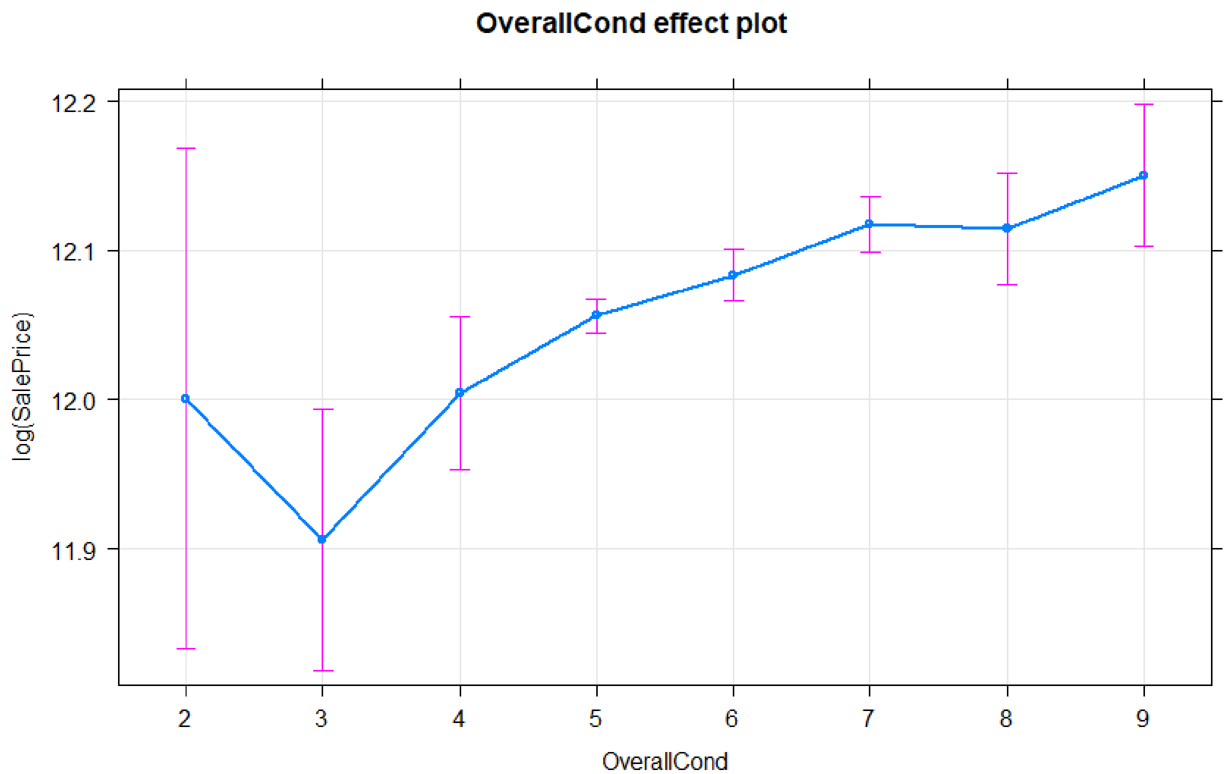

OverallCond effect plot
log(SalePrice)
12.2
12.1
12.0
11.9
2
3
4
5
6
7
8
9
OverallCond

Figure 32

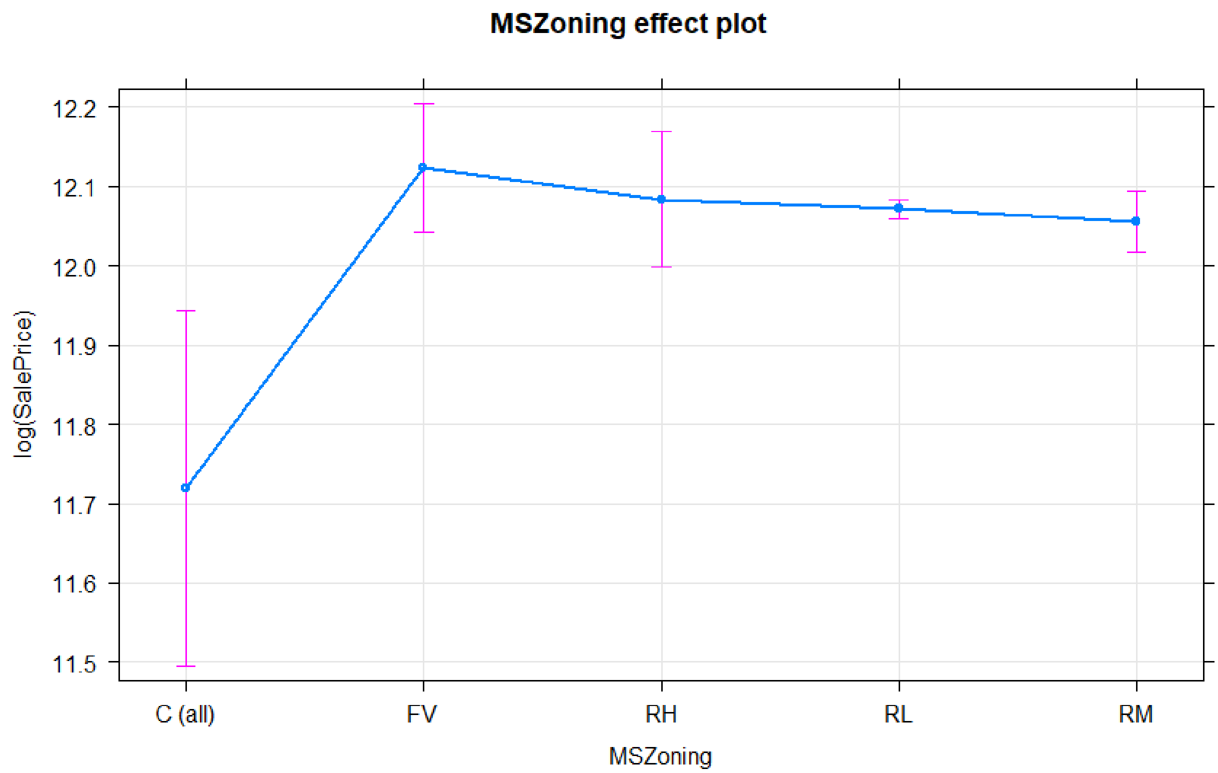

MSZoning effect plot
log(SalePrice)
12.2
12.1
12.0
11.9
11.8
11.7
11.6
11.5
C (all)
FV
RH
RL
RM
MSZoning

Figure 33

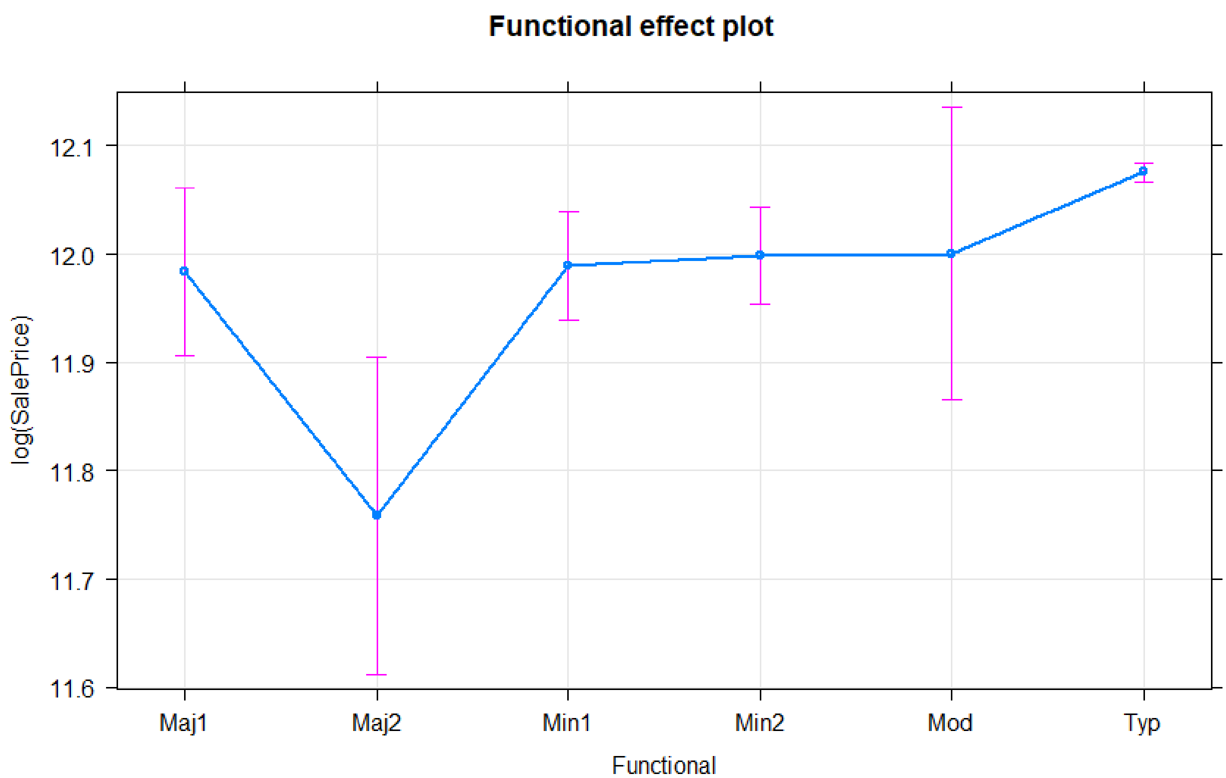

Functional effect plot
12.1
12.0
11.9
11.8
11.7
11.6
log(SalePrice)
Maj1
Maj2
Min1
Min2
Mod
Typ
Functional

Figure 34


SaleCondition effect plot
12.2
12.1
12.0
11.9
log(SalePrice)
Abnorml
AdjLand
Alloca
Family
Normal
Partial
SaleCondition

Figure 35

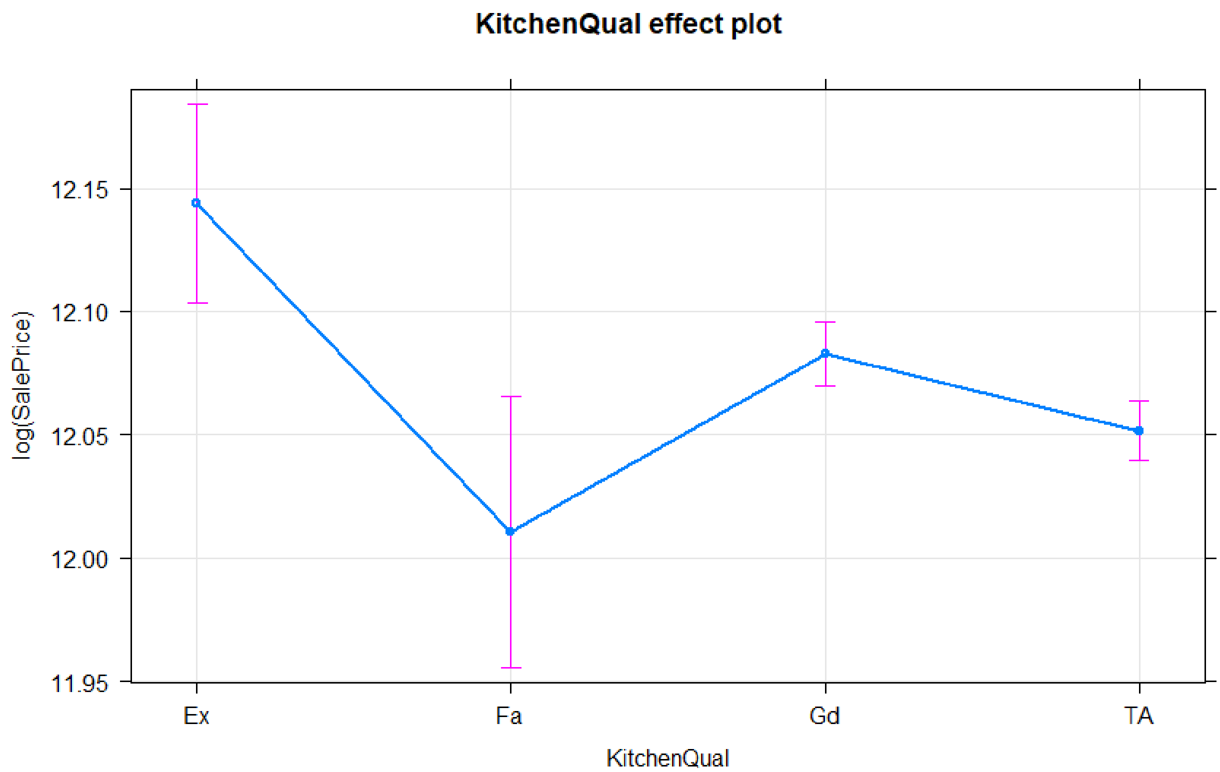

KitchenQual effect plot
log(SalePrice)
12.15
12.10
12.05
12.00
11.95
Ex
Fa
Gd
TA
KitchenQual

Figure 36