Analysis on Value of Residential Homes in Ames, Iowa

Lingzhi Du, Gongting Peng, Xiaowen Zhang, Yue Lan

Contents

1. Data Analysis

1.1 Data Description

The data is a collection of basic housing information and sales record residential homes in Ames, Iowa sold in 2006-2010. It contains 1460 observations with 81 variables, including 23 nominal, 23 ordinal, 14 discrete, 19 continuous variables, a house ID, and the sale price of each house. The 79 explanatory variables can be summarized as follows.

	Explanatory Variables	Data Type
Overall	OverallQual, OverallCond	Ordinal
Type/Style	MSSubClass, BldgType, HouseStyle	Nominal
Location	MSZoing, Neighborhood, Condition1, Condition2	Nominal
Lot	LotFrontage, LotArea	Continuous
	LotShape, LandSlope	Ordinal
	Street, Alley, LandContour, LotConfig	Nominal
Utilities	Utilities, HeatingQC, Electrical	Ordinal
	Heating, CentralAir	Nominal
Area	1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea	Continuous
Year	YearBuilt, YearRemodAdd	Discrete
Exterior	RoofStyle, RoofMatl, Exterior1st, Exterior2nd,	
	MasVnrType, Foundation	Nominal
	MasVnrArea	Continuous
	ExterQual, ExterCond	Ordinal
Interior	BsmtFullBath, BsmtHalfBath, FullBath, HalfBath,	
	Kitchen, Bedroom, Fireplaces, TotRmsAbvGrd	Discrete
	Functional, KitchenQual, FireplaceQu	Ordinal
Basement	BsmtQual, BsmtCond, BsmtExposure,	
	BsmtFinType1, BsmtFinType2	Ordinal
	BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF	Continuous
Garage	GarageType	Nominal
	GarageYrBlt, GarageCars	Discrete
	GarageFinish, PavedDrive, GarageQual, GarageCond	Ordinal
	GarageArea	Continuous
Others	OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch,	
	WoodDeckSF, PoolArea	Continuous
	PoolQC, Fence	Ordinal
Miscellaneous	MiscFeature	Nominal
	MiscVal	Continuous
Sale	MoSold, YrSold	Discrete
	SaleType, SaleCondition	Nominal

The variables measure the physical attributes of the properties that a potential home buyer is concerned with. A majority of them are categorical variables that generally identify the type (nominal variables) of certain dwelling feature or evaluate the quality and condition (ordinal variables) of a specific aspect of the house. Discrete variables are often counts of particular items such as rooms or records of dates. Continuous variables typically quantify the area and dimensions for different attributes. Some of the explanatory variables highly correlates with another in nature by definition.

1.2 Data Pre-processing

1.2.1 Data Types

In accordance with data dictionary, nominal variables are loaded as factors, ordinal variables are loaded as ordered factors, and discrete and continuous varibles are loaded as vectors of integers.

Categorical data will be translated into dummy variables for each category when fitting models.

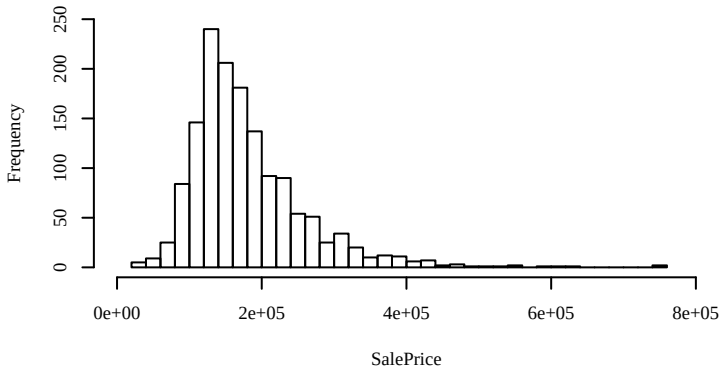
1.2.2 Response Variable

The response variable we focus on is the sale prices of houses. In the dataset, the values of responses run across different orders of magnitude, ranging from 34,900 to 755,000. The distribution of sale prices is right-skewed with a mean of 180,921.

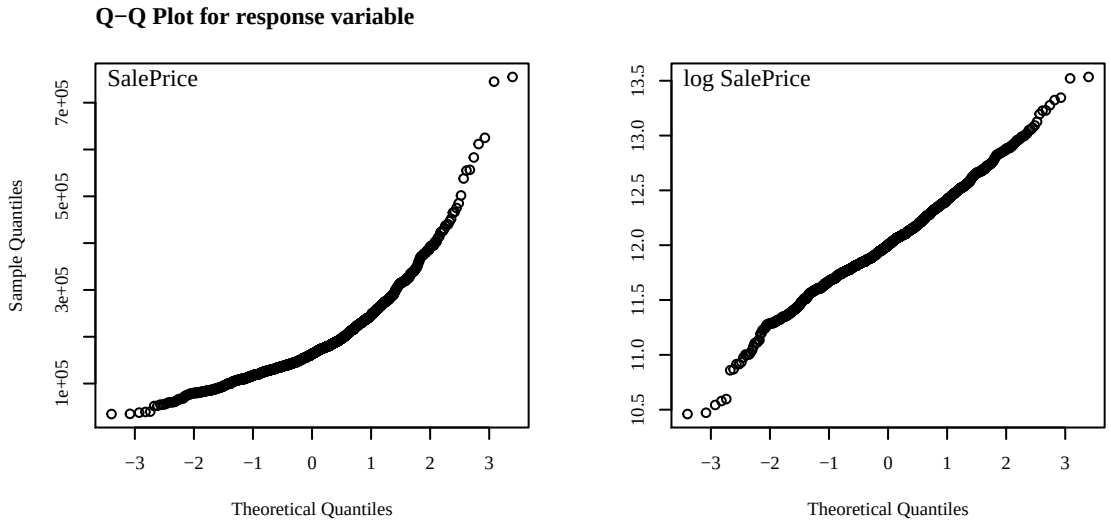
`Summary of SalePrice:`

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34900  129975  163000  180921  214000  755000
```

`Histogram of SalePrice:`



The skewness and variability of data can be reduced by taking a log tranformation. After transformed into its logrithm, we can see that `SalePrice` displays a close pattern towards normal distribution with a quantile-quantile plot.



The nice feature of conformation to normality can help improve the fit of a linear model and correct violations of model assumptions such as nonconstant error variance. Therefore we will use the logrithm of sales price as the measurement of response variable when building up the model. A formal test of normality will be performed when diagnosing the model.

1.2.3 Missing Values

- `NA` values

  A total of 19 variables have `NA` values in the data set, 15 of which are categorical data using `NA` to denote 'no such feature' and 4 have missing observations.

- `0`s that can be interpreted as missing values

  For some numeric variables, a value of `0` may be an indicator of 'not applicable'. For example, a `GarageArea` of 0 is equivalent to not having a garage. If we are examining the effect of garage spaciousness on housing prices, the 0s can lead to an exaggaration of the corresponding parameter. We will treat these irrelevant 0s as missing values as well.

  To deal with missing values in categorical data, a new level of `NA` is added to each of the factors. To deal with missing values in numeric values, we are adopting the mean imputation method. In this way, both sample size and sample mean maintain the same, and there is guaranteed to be no correlation between the imputed values and other variables. However, the imputation reduces the variability of the variables themselves as well as their covariances with other variables. The under-statement of variance may result in an under-estimate of p-value of the corresponding parameter estimator.

1.2.4 Dependence and Multi-colinearality

- Overall evaluation variables

  In the dataset there are two variables `OverallQual` and `OverallCond` that evaluate the quality and condition of the houses respectively in general. These variables are derived from the evaluation of other features and thus contain little information themselves. In addition, they are hard to interpret from an explanatory perspective. Therefore we will not include them in the explanatory model.

- Exact multi-colinearality

  According to data dictionary, there are perfect linear relationships among certain variables. The above grade (ground) living area square feet `GrlivArea` is the sum of first floor square feet `X1stFlrSF`, second floor square feet `X2ndFlrSF`, and low quality finished square feet (all floors) `LowQualFinSF`; and the total square feet of basement area `TotalBsmtSF` is the sum of type 1 finished square feet `BsmtFinSF1`, type 2 finished square feet `BsmtFinSF2`, and unfinished square feet of basement area `BsmtUnfSF`. The data itself also gives that $Cor(GrlivArea, X1stFlrSF + X2ndFlrSF + LowQualFinSF) = 1$ and $Cor(TotalBsmtSF, BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF) = 1$.

  To avoid exact multi-colinearality issue, we'll remove `GrLivArea` and `TotalBsmtSF` from the dataset.

- Other issues related to multi-colinearality

  Most of the categorical data in the data set are highly imbalanced across different levels. It is often the case that a dominant category takes up more than 90% percent of the data and the scarse categories take up very small percetages. Here are some examples of extreme cases.

```
   RoofMatl     Heating      Street      Utilities
CompShg:1434   Floor:   1   Grvl:   6   AllPub:1459
Tar&Grv:  11   GasA :1428   Pave:1454   NoSeWa:   1
WdShngl:   6   GasW :  18
WdShake:   5   Grav :   7
ClyTile:   1   OthW :   2
Membran:   1   Wall :   4
(Other):   2
```

  As these factors are reorganized into dummy variables, the vectors will be composed of generally all 1s for dominant categories and all 0s for minority categories, thus highly correlated with each other as well as with the intercept term. It will give rise to multi-colinearality issue.

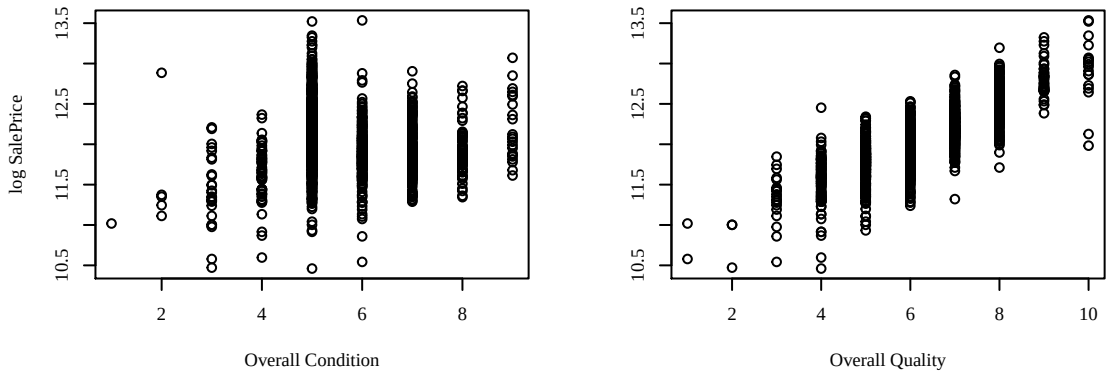
  Meanwhile, the sparseness of the variables can boost the variance of corresponding parameter estimators and hence reduce the significance level. It is very likely that these variables will be eliminated from the model during the variable selection process. Therefore, we will keep them for now and revisit the issue when evaluating the final model.

1.3 Exploratory Analysis

- Overall Evaluations:

  As aforementioned, there are two variables that evaluate the houses' overall qualities and conditions. They are directly associated with the sale prices.

Overall Evaluations

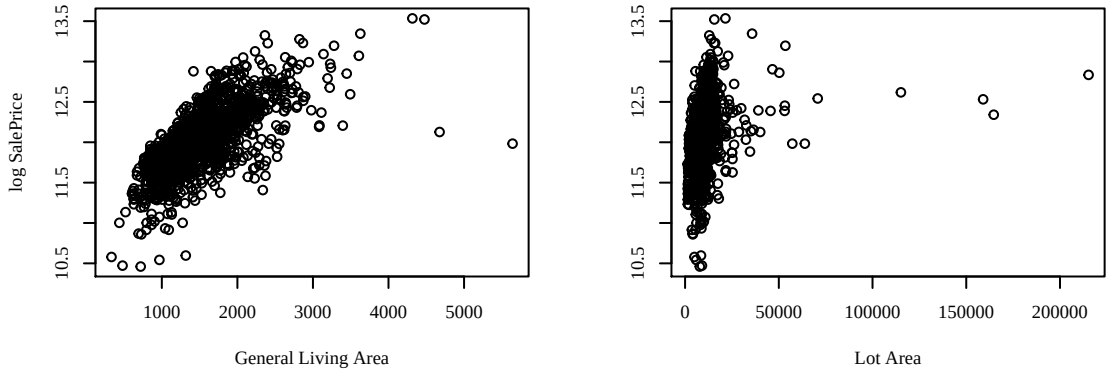


We can see from the above graph that there is a positive linear relationship between housing prices and each of the evaluation scores. The price in log scale generally goes up as the evaluation scores go up. Overall quality of a property is a stronger indicator of sale prices since the clearer trend suggests a higher correlation ($corr = 0.8171844$).

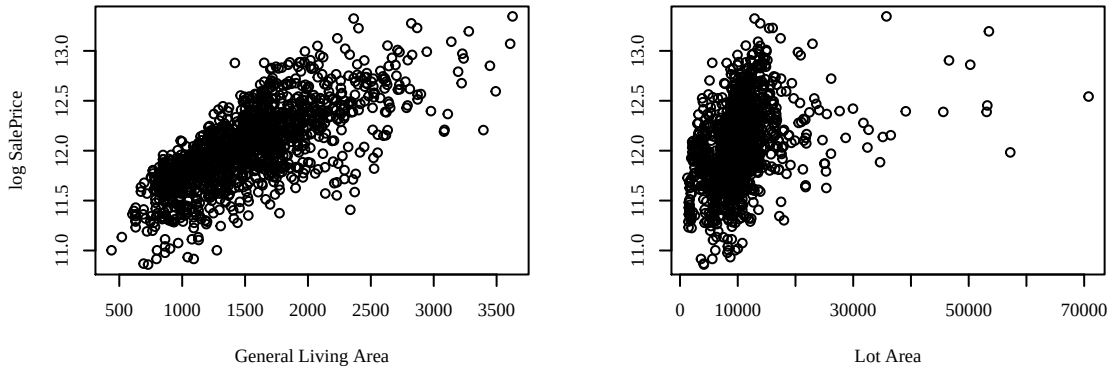
- Impact of Area

  Intuitively, area plays a key role in determining the price of property. Here is a plot that shows the relationships between sale prices and the main area statistics of houses.

Impact of Area



Impact of Area (outliers removed)

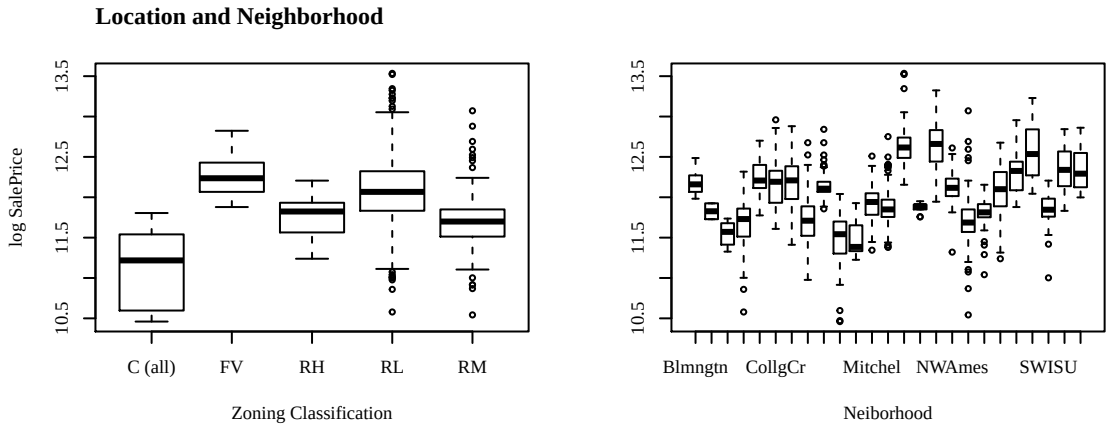


With the plot, it is obvious that there are 4 outliers with significantly large areas and 5 outliers with implausible low sale prices. We can have a better view of the relationship between variables after removing these outliers.

We will use a formal statistical method to detect the outliers later during the model building process.

- Location and Neighborhood

  Another intuition is that location is essential for the pricing of propertities. From the plot we can see that one zone or neighborhood can distinguish from another in terms of average housing prices.



2. Explanatory Model

2.1 Model Building

2.1.1 Variable Selection with LASSO

Since post-selection inference for regression coefficients that conditions on model selection with LASSO has been proved to be valid, we will first adopt the LASSO model to preliminarily eliminate explanatory variables.

To reach for the best tuning parameter λ, we randomly split the data set into training set against test set and then apply a 5-fold cross-validation recursively for 10 times. Each time we get a best λ corresponding to that random partition of training set and test set. The final best λ is derived by taking the mean of the 10 λs.

```
> $best \lambda = 0.0058872$
```

Note that the dataset after pre-processing procedures has 97 variables, which extend to 261 variables after re-organizing the categorical variables. With implementation of LASSO, the number of variables has shrinked significantly to 54 variables.

In the latter part of the report, we will refer to the model with 54 explanatory variables selected using LASSO as the full model.

2.1.2 Stepwise Feature Selection with ANOVA F-test and BIC

- ANOVA F-test

  As we inspect the full model, we can observe the parameters with large p-value from the summary of the full model.

Variables	Estimates	Std. Error	t value	p-value	siginificance
SaleTypeCon	1.043e-01	1.042e-01	1.001	0.316815	
OpenPorchSF	1.248e-04	6.394e-05	1.952	0.051141	.
GarageArea	1.240e-05	3.254e-05	0.381	0.703219	
HeatingQC.L	6.294e-02	3.562e-02	1.767	0.077422	.
HeatingQC.Q	-3.879e-03	2.165e-02	-0.179	0.857786	

Variables	Estimates	Std. Error	t value	p-value	siginificance
BsmtExposure.L	3.486e-03	1.148e-02	0.304	0.761539	
ExterCond.L	5.663e-02	3.568e-02	1.587	0.112717	
MasVnrArea	4.705e-05	2.556e-05	1.840	0.065912	.
NeighborhoodVeenker	7.273e-02	4.550e-02	1.599	0.110111	
StreetPave	9.996e-02	6.142e-02	1.628	0.103829	
MSZoningFV	6.853e-02	3.867e-02	1.772	0.076610	.
KitchenQual.Q	3.457e-02	1.503e-02	2.300	0.021584	*
HeatingGasW	7.993e-02	3.675e-02	2.175	0.029813	*
RoofMatlMembran	3.467e-01	1.473e-01	2.354	0.018698	*

With the identification of variables that have large or relatively large p-values, we will be performing ANOVA F-test to check these variables one by one to see if each of them has a significant impact on our model at a significance level of $\alpha = 0.01$. We will start with testing the variable with the largest p-value in the current model, and consider removing it if the significance criteria is not met.

```
Analysis of Variance Table

Model 1: y ~ MSSubClass50 + MSSubClass60 + MSSubClass70 + MSZoningFV +
    MSZoningRL + LotArea + StreetPave + LandContourHLS + LotConfigCulDSac +
    NeighborhoodBrkSide + NeighborhoodClearCr + NeighborhoodCrawfor +
    NeighborhoodNoRidge + NeighborhoodNridgHt + NeighborhoodSomerst +
    NeighborhoodStoneBr + NeighborhoodVeenker + YearRemodAdd +
    RoofMatlMembran + RoofMatlWdShngl + MasVnrArea + ExterQual.L +
    ExterCond.L + FoundationPConc + BsmtExposure.L + BsmtFinSF1 +
    HeatingGasW + HeatingQC.L + HeatingQC.Q + CentralAirTRUE +
    X1stFlrSF + X2ndFlrSF + BsmtFullBath + FullBath + HalfBath +
    KitchenQual.L + KitchenQual.Q + TotRmsAbvGrd + Functional.L +
    Fireplaces + GarageTypeAttchd + GarageCars + GarageArea +
    PavedDrive.L + WoodDeckSF + OpenPorchSF + ScreenPorch + PoolArea +
    SaleTypeCon + SaleTypeNew + SaleConditionNormal + ConditionNorm +
    ExteriorBrkFace
Model 2: y ~ MSSubClass50 + MSSubClass60 + MSSubClass70 + MSZoningFV +
    MSZoningRL + LotArea + StreetPave + LandContourHLS + LotConfigCulDSac +
    NeighborhoodBrkSide + NeighborhoodClearCr + NeighborhoodCrawfor +
    NeighborhoodNoRidge + NeighborhoodNridgHt + NeighborhoodSomerst +
    NeighborhoodStoneBr + NeighborhoodVeenker + YearRemodAdd +
    RoofMatlMembran + RoofMatlWdShngl + MasVnrArea + ExterQual.L +
    ExterCond.L + FoundationPConc + BsmtExposure.L + BsmtFinSF1 +
    HeatingGasW + HeatingQC.L + CentralAirTRUE + X1stFlrSF +
    X2ndFlrSF + BsmtFullBath + FullBath + HalfBath + KitchenQual.L +
    KitchenQual.Q + TotRmsAbvGrd + Functional.L + Fireplaces +
    GarageTypeAttchd + GarageCars + GarageArea + PavedDrive.L +
    WoodDeckSF + OpenPorchSF + ScreenPorch + PoolArea + SaleTypeCon +
    SaleTypeNew + SaleConditionNormal + ConditionNorm + ExteriorBrkFace
  Res.Df    RSS Df   Sum of Sq      F Pr(>F)
1   1406 28.275
2   1407 28.276 -1 -0.00064599 0.0321 0.8578
```

As we can see in the summary above, the p-value for ANOVA F-test is 0.8578, which is greater than the set significent level of 0.01. There is not enough evidence to prove that `HeatingQC.Q` has an effect on `SalePrice`. Therefore, we remove this parameter from our full model and test our reduced model with the same approach.

After iterating this process for 13 times, we derive our final "best" model with every parameter having a significant impact on our regression model.

```
Call:
lm(formula = y ~ MSSubClass50 + MSSubClass60 + MSSubClass70 +
    MSZoningRL + LotArea + LandContourHLS + LotConfigCulDSac +
    NeighborhoodBrkSide + NeighborhoodClearCr + NeighborhoodCrawfor +
    NeighborhoodNoRidge + NeighborhoodNridgHt + NeighborhoodSomerst +
    NeighborhoodStoneBr + YearRemodAdd + RoofMatlWdShngl + ExterQual.L +
    FoundationPConc + BsmtFinSF1 + HeatingQC.L + CentralAirTRUE +
    X1stFlrSF + X2ndFlrSF + BsmtFullBath + FullBath + HalfBath +
    KitchenQual.L + TotRmsAbvGrd + Functional.L + Fireplaces +
    GarageTypeAttchd + GarageCars + PavedDrive.L + WoodDeckSF +
    ScreenPorch + PoolArea + SaleTypeNew + SaleConditionNormal +
    ConditionNorm + ExteriorBrkFace, data = subsetHousing)

Residuals:
     Min       1Q   Median       3Q      Max
-1.30294 -0.06358  0.00986  0.07749  0.47711

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.454e+00  6.848e-01   5.043 5.17e-07 ***
MSSubClass50        1.161e-01  1.629e-02   7.129 1.60e-12 ***
MSSubClass60        6.526e-02  1.598e-02   4.085 4.65e-05 ***
MSSubClass70        1.451e-01  2.231e-02   6.502 1.09e-10 ***
MSZoningRL          5.033e-02  1.132e-02   4.446 9.44e-06 ***
LotArea             1.578e-06  4.355e-07   3.624 0.000300 ***
LandContourHLS      5.894e-02  2.180e-02   2.704 0.006942 **
LotConfigCulDSac    4.915e-02  1.600e-02   3.072 0.002165 **
NeighborhoodBrkSide 6.987e-02  2.095e-02   3.334 0.000877 ***
NeighborhoodClearCr 8.967e-02  2.913e-02   3.078 0.002122 **
NeighborhoodCrawfor 1.346e-01  2.320e-02   5.802 8.06e-09 ***
NeighborhoodNoRidge 1.190e-01  2.554e-02   4.660 3.46e-06 ***
NeighborhoodNridgHt 1.601e-01  1.995e-02   8.025 2.11e-15 ***
NeighborhoodSomerst 1.422e-01  1.952e-02   7.286 5.29e-13 ***
NeighborhoodStoneBr 1.865e-01  3.074e-02   6.065 1.69e-09 ***
YearRemodAdd        1.980e-03  2.777e-04   7.131 1.58e-12 ***
RoofMatlWdShngl     1.981e-01  6.070e-02   3.263 0.001128 **
ExterQual.L         1.630e-01  3.527e-02   4.622 4.14e-06 ***
FoundationPConc     3.294e-02  1.116e-02   2.952 0.003213 **
BsmtFinSF1          4.224e-05  1.223e-05   3.455 0.000567 ***
HeatingQC.L         5.595e-02  1.645e-02   3.401 0.000691 ***
CentralAirTRUE      1.302e-01  1.764e-02   7.379 2.71e-13 ***
X1stFlrSF           2.064e-04  1.789e-05  11.536  < 2e-16 ***
X2ndFlrSF           2.265e-04  2.877e-05   7.874 6.75e-15 ***
BsmtFullBath        4.904e-02  9.695e-03   5.059 4.77e-07 ***
FullBath            6.990e-02  1.023e-02   6.830 1.26e-11 ***
HalfBath            6.249e-02  1.022e-02   6.116 1.24e-09 ***
KitchenQual.L       9.448e-02  2.061e-02   4.584 4.97e-06 ***
TotRmsAbvGrd        2.144e-02  3.861e-03   5.553 3.34e-08 ***
Functional.L        3.343e-01  3.801e-02   8.796  < 2e-16 ***
Fireplaces          4.438e-02  7.259e-03   6.114 1.25e-09 ***
```

```
GarageTypeAttchd    3.820e-02  9.559e-03   3.996 6.76e-05 ***
GarageCars          7.116e-02  6.942e-03  10.249  < 2e-16 ***
PavedDrive.L        5.025e-02  1.219e-02   4.122 3.98e-05 ***
WoodDeckSF          1.064e-04  3.326e-05   3.198 0.001413 **
ScreenPorch         3.616e-04  6.991e-05   5.173 2.64e-07 ***
PoolArea            5.187e-03  6.811e-04   7.615 4.80e-14 ***
SaleTypeNew         1.114e-01  2.007e-02   5.551 3.38e-08 ***
SaleConditionNormal 8.775e-02  1.314e-02   6.677 3.49e-11 ***
ConditionNorm       1.410e-01  3.846e-02   3.666 0.000255 ***
ExteriorBrkFace     6.152e-02  2.142e-02   2.872 0.004139 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.143 on 1419 degrees of freedom
Multiple R-squared:  0.8754,  Adjusted R-squared:  0.8719
F-statistic: 249.3 on 40 and 1419 DF,  p-value: < 2.2e-16
```

By doing ANOVA F-test, we have further eliminated 13 variables from the full model. We are going to use an alternative approach, stepwise BIC, to find the "best" model and hopefully we can get the same "best" result.

- Stepwise BIC

```
Start:  AIC=-5365.08
y ~ MSSubClass50 + MSSubClass60 + MSSubClass70 + MSZoningFV +
    MSZoningRL + LotArea + StreetPave + LandContourHLS + LotConfigCulDSac +
    NeighborhoodBrkSide + NeighborhoodClearCr + NeighborhoodCrawfor +
    NeighborhoodNoRidge + NeighborhoodNridgHt + NeighborhoodSomerst +
    NeighborhoodStoneBr + NeighborhoodVeenker + YearRemodAdd +
    RoofMatlMembran + RoofMatlWdShngl + MasVnrArea + ExterQual.L +
    ExterCond.L + FoundationPConc + BsmtExposure.L + BsmtFinSF1 +
    HeatingGasW + HeatingQC.L + HeatingQC.Q + CentralAirTRUE +
    X1stFlrSF + X2ndFlrSF + BsmtFullBath + FullBath + HalfBath +
    KitchenQual.L + KitchenQual.Q + TotRmsAbvGrd + Functional.L +
    Fireplaces + GarageTypeAttchd + GarageCars + GarageArea +
    PavedDrive.L + WoodDeckSF + OpenPorchSF + ScreenPorch + PoolArea +
    SaleTypeCon + SaleTypeNew + SaleConditionNormal + ConditionNorm +
    ExteriorBrkFace

                       Df Sum of Sq    RSS     AIC
- HeatingQC.Q           1   0.00065 28.276 -5372.3
- BsmtExposure.L        1   0.00185 28.277 -5372.3
- GarageArea            1   0.00292 28.278 -5372.2
- SaleTypeCon           1   0.02017 28.296 -5371.3
- ExterCond.L           1   0.05066 28.326 -5369.8
- NeighborhoodVeenker   1   0.05140 28.327 -5369.7
- StreetPave            1   0.05328 28.329 -5369.6
- HeatingQC.L           1   0.06280 28.338 -5369.1
- MSZoningFV            1   0.06315 28.338 -5369.1
- MasVnrArea            1   0.06812 28.343 -5368.9
- OpenPorchSF           1   0.07662 28.352 -5368.4
- HeatingGasW           1   0.08506 28.360 -5368.0
- RoofMatlMembran       1   0.11472 28.390 -5366.5
- NeighborhoodClearCr   1   0.12901 28.404 -5365.7
- KitchenQual.Q         1   0.12980 28.405 -5365.7
```

```
<none>                              28.275 -5365.1
- NeighborhoodSomerst  1   0.14230 28.418 -5365.0
- BsmtFinSF1           1   0.15724 28.433 -5364.3
- LotConfigCulDSac     1   0.16099 28.436 -5364.1
- LandContourHLS       1   0.16293 28.438 -5364.0
- ExteriorBrkFace      1   0.16541 28.441 -5363.8
- RoofMatlWdShngl      1   0.19320 28.469 -5362.4
- NeighborhoodBrkSide  1   0.21358 28.489 -5361.4
- FoundationPConc      1   0.21495 28.490 -5361.3
- WoodDeckSF           1   0.23126 28.507 -5360.5
- LotArea              1   0.24242 28.518 -5359.9
- MSSubClass60         1   0.28638 28.562 -5357.7
- ExterQual.L          1   0.29387 28.569 -5357.3
- ConditionNorm        1   0.30257 28.578 -5356.8
- KitchenQual.L        1   0.33897 28.614 -5355.0
- GarageTypeAttchd     1   0.35705 28.632 -5354.0
- PavedDrive.L         1   0.37982 28.655 -5352.9
- NeighborhoodNoRidge  1   0.42061 28.696 -5350.8
- ScreenPorch          1   0.48793 28.763 -5347.4
- BsmtFullBath         1   0.50731 28.783 -5346.4
- MSZoningRL           1   0.51980 28.795 -5345.8
- HalfBath             1   0.55778 28.833 -5343.8
- SaleTypeNew          1   0.59282 28.868 -5342.1
- Fireplaces           1   0.66613 28.942 -5338.4
- TotRmsAbvGrd         1   0.66790 28.943 -5338.3
- NeighborhoodCrawfor  1   0.70985 28.985 -5336.2
- NeighborhoodStoneBr  1   0.73221 29.008 -5335.0
- MSSubClass70         1   0.82855 29.104 -5330.2
- SaleConditionNormal  1   0.87698 29.152 -5327.8
- FullBath             1   0.92307 29.198 -5325.5
- MSSubClass50         1   0.94245 29.218 -5324.5
- NeighborhoodNridgHt  1   0.95411 29.230 -5323.9
- YearRemodAdd         1   0.99499 29.270 -5321.9
- X2ndFlrSF            1   1.03839 29.314 -5319.7
- PoolArea             1   1.10927 29.385 -5316.2
- CentralAirTRUE       1   1.16588 29.441 -5313.4
- Functional.L         1   1.53765 29.813 -5295.1
- GarageCars           1   1.56183 29.837 -5293.9
- X1stFlrSF            1   2.17804 30.453 -5264.0

Error in eval(predvars, data, env): object 'MSZoningFV' not found
```

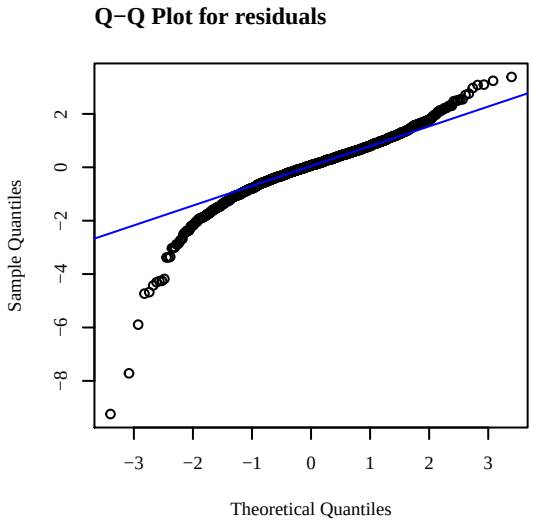
As we can see from the result above, the model with the minimum AIC value of -5422.88 is eliminating the exact same 13 variables as the previous model is, which is the same as "model13" aka our "best" model in the anova F-test part.

Therefore, we got the same "best" model from ANOVA F-test and stepwise BIC.

2.1.3 Model Justification and Verification

- Test for Normality

  Below is a plot of the standardized residuals:



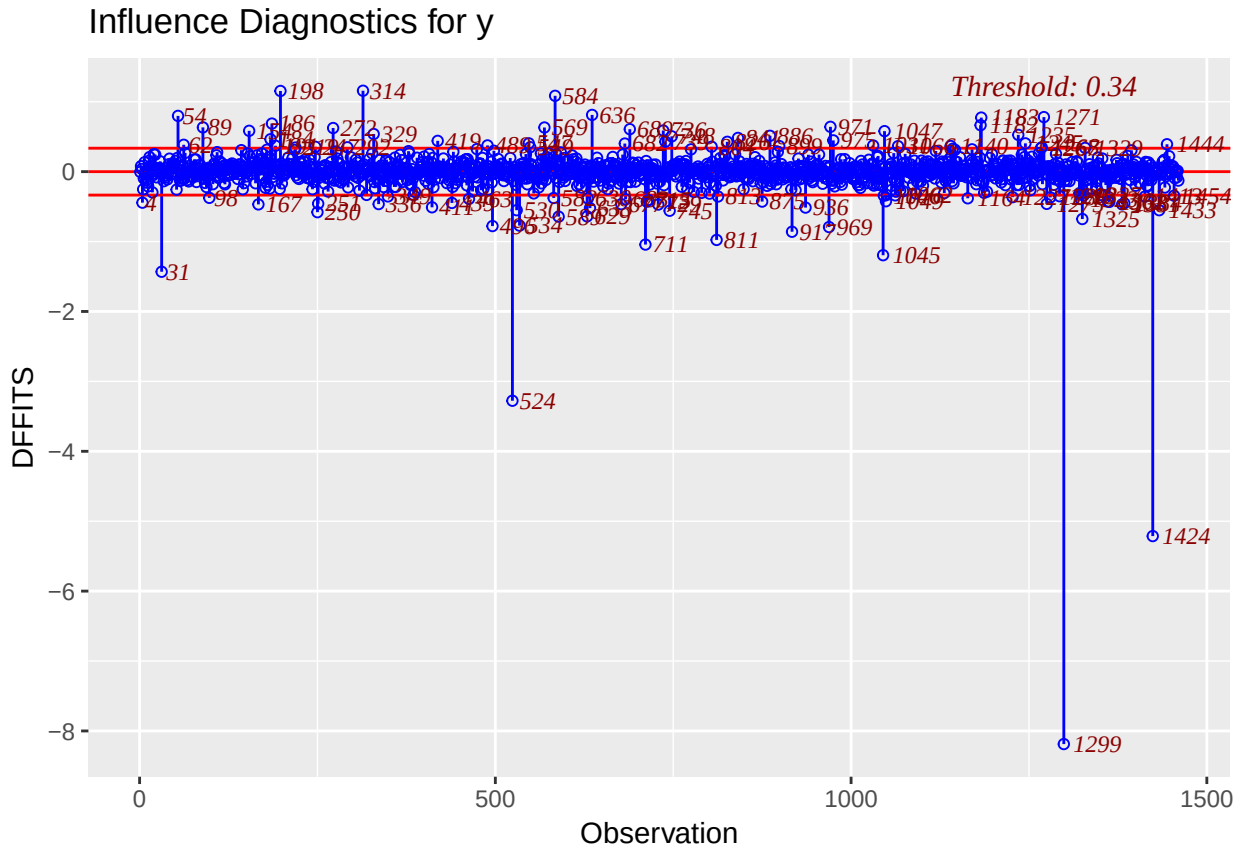
As shown in the graph, the residuals are approximately normal around its mean. However, the distribution of residuals seems to have heavy tails on both ends.

To give a general conclusion, we are performing Kolmogorov-Smirnov to formally test the conformation of normality quantitatively. We are running the Kolmogorov-Smirnov test for 100 times and taking the mean of 100 results as the ks score.

After running the tests, we get a mean p-value of 0.002391308, which is below our chosen significance level. Therefore we reject the null hypothesis that the residuals are normally distributed.

Thus, we conclude that even though the residuals looks normally distributed in some ways, we fail to meet our assumptions according to the Kolmogorov-Smirnov test.

- Influential Points and Outliers:



From the plot above, we can observe that there are many influential points and outliers that are shifting our model and causing the violation of normality assumption. These data points should be removed.

We use dffits to remove outliers. We set the dffits value bar to be $2 * sqrt(p/n)$, where p is the number of parameter in the dataset and n is number of observations in the dataset. In our case, the quantity of this bar is 0.3310424, so we remove all the observations who's dffits value exceeds this bar and fit a new full model with the updated dataset.

- Updated ANOVA F-test

  Since we are updating the data and full model, we need to re-fit our model.

Variables	Estimates	Std. Error	t value	p-value	siginificance
NeighborhoodClearCr	4.555e-02	2.542e-02	1.792	0.073407	.
RoofMatlWdShngl	1.387e-01	7.446e-02	1.863	0.062738	.
PoolArea	4.486e-03	1.835e-03	2.444	0.014658	*

After identifying the variables with relatively large p-value, we will re-run ANOVA F-test to check these variables one by one to see if each of them has a significant impact on our model at a significance level of $\alpha = 0.01$.

```
Analysis of Variance Table

Model 1: y ~ MSSubClass50 + MSSubClass60 + MSSubClass70 + MSZoningRL +
    LotArea + LandContourHLS + LotConfigCulDSac + NeighborhoodBrkSide +
    NeighborhoodClearCr + NeighborhoodCrawfor + NeighborhoodNoRidge +
```

```
    NeighborhoodNridgHt + NeighborhoodSomerst + NeighborhoodStoneBr +
    YearRemodAdd + RoofMatlWdShngl + ExterQual.L + FoundationPConc +
    BsmtFinSF1 + HeatingQC.L + CentralAirTRUE + X1stFlrSF + X2ndFlrSF +
    BsmtFullBath + FullBath + HalfBath + KitchenQual.L + TotRmsAbvGrd +
    Functional.L + Fireplaces + GarageTypeAttchd + GarageCars +
    PavedDrive.L + WoodDeckSF + ScreenPorch + PoolArea + SaleTypeNew +
    SaleConditionNormal + ConditionNorm + ExteriorBrkFace
Model 2: y ~ MSSubClass50 + MSSubClass60 + MSSubClass70 + MSZoningRL +
    LotArea + LandContourHLS + LotConfigCulDSac + NeighborhoodBrkSide +
    NeighborhoodCrawfor + NeighborhoodNoRidge + NeighborhoodNridgHt +
    NeighborhoodSomerst + NeighborhoodStoneBr + YearRemodAdd +
    RoofMatlWdShngl + ExterQual.L + FoundationPConc + BsmtFinSF1 +
    HeatingQC.L + CentralAirTRUE + X1stFlrSF + X2ndFlrSF + BsmtFullBath +
    FullBath + HalfBath + KitchenQual.L + TotRmsAbvGrd + Functional.L +
    Fireplaces + GarageTypeAttchd + GarageCars + PavedDrive.L +
    WoodDeckSF + ScreenPorch + PoolArea + SaleTypeNew + SaleConditionNormal +
    ConditionNorm + ExteriorBrkFace
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   1315 13.064
2   1316 13.096 -1 -0.031893 3.2103 0.07341 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see in the summary above, the p-value for anova F-test is 0.07341, which is greater than the chosen significent level of 0.01. Thus, there is not enough evidence to prove that `NeighborhoodClearCr` has an effect on SalePrice. Therefore, we will remove the variable from our new full model and test our new reduced model with the same approach.

After iterating this process for 3 times, we identify our new best model with every parameter having significent impact on our new regression model.

```
Call:
lm(formula = y ~ MSSubClass50 + MSSubClass60 + MSSubClass70 +
    MSZoningRL + LotArea + LandContourHLS + LotConfigCulDSac +
    NeighborhoodBrkSide + NeighborhoodCrawfor + NeighborhoodNoRidge +
    NeighborhoodNridgHt + NeighborhoodSomerst + NeighborhoodStoneBr +
    YearRemodAdd + ExterQual.L + FoundationPConc + BsmtFinSF1 +
    HeatingQC.L + CentralAirTRUE + X1stFlrSF + X2ndFlrSF + BsmtFullBath +
    FullBath + HalfBath + KitchenQual.L + TotRmsAbvGrd + Functional.L +
    Fireplaces + GarageTypeAttchd + GarageCars + PavedDrive.L +
    WoodDeckSF + ScreenPorch + SaleTypeNew + SaleConditionNormal +
    ConditionNorm + ExteriorBrkFace, data = dffits_data_bs)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46180 -0.05783  0.00475  0.06694  0.28898

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.860e+00  4.058e-01  16.906  < 2e-16 ***
MSSubClass50       1.132e-01  1.237e-02   9.150  < 2e-16 ***
MSSubClass60       1.163e-01  1.195e-02   9.729  < 2e-16 ***
MSSubClass70       1.756e-01  1.734e-02  10.128  < 2e-16 ***
MSZoningRL         4.970e-02  8.500e-03   5.847 6.32e-09 ***
```

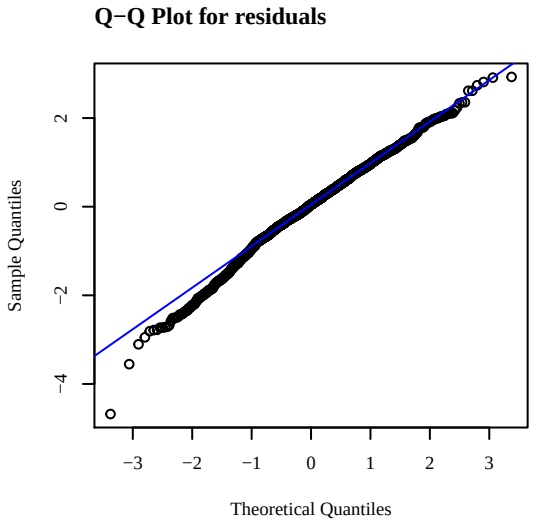
```
LotArea             4.404e-06  5.606e-07   7.856 8.20e-15 ***
LandContourHLS      5.424e-02  1.713e-02   3.167 0.001576 **
LotConfigCulDSac    3.969e-02  1.180e-02   3.364 0.000790 ***
NeighborhoodBrkSide 7.575e-02  1.627e-02   4.655 3.56e-06 ***
NeighborhoodCrawfor 1.255e-01  1.766e-02   7.106 1.95e-12 ***
NeighborhoodNoRidge 1.024e-01  1.866e-02   5.490 4.81e-08 ***
NeighborhoodNridgHt 1.414e-01  1.446e-02   9.777  < 2e-16 ***
NeighborhoodSomerst 1.488e-01  1.411e-02  10.551  < 2e-16 ***
NeighborhoodStoneBr 1.646e-01  2.341e-02   7.033 3.25e-12 ***
YearRemodAdd        1.783e-03  2.052e-04   8.685  < 2e-16 ***
ExterQual.L         1.684e-01  2.624e-02   6.418 1.92e-10 ***
FoundationPConc     3.460e-02  8.176e-03   4.232 2.48e-05 ***
BsmtFinSF1          7.672e-05  9.260e-06   8.285 2.88e-16 ***
HeatingQC.L         4.642e-02  1.201e-02   3.867 0.000116 ***
CentralAirTRUE      1.367e-01  1.395e-02   9.797  < 2e-16 ***
X1stFlrSF           2.631e-04  1.378e-05  19.090  < 2e-16 ***
X2ndFlrSF           1.972e-04  2.222e-05   8.875  < 2e-16 ***
BsmtFullBath        2.901e-02  7.260e-03   3.996 6.79e-05 ***
FullBath            6.242e-02  7.726e-03   8.080 1.45e-15 ***
HalfBath            5.710e-02  7.577e-03   7.536 8.99e-14 ***
KitchenQual.L       1.042e-01  1.518e-02   6.865 1.02e-11 ***
TotRmsAbvGrd        1.348e-02  2.963e-03   4.550 5.86e-06 ***
Functional.L        3.241e-01  2.973e-02  10.902  < 2e-16 ***
Fireplaces          3.529e-02  5.296e-03   6.663 3.94e-11 ***
GarageTypeAttchd    2.532e-02  6.924e-03   3.657 0.000266 ***
GarageCars          5.911e-02  5.276e-03  11.204  < 2e-16 ***
PavedDrive.L        5.879e-02  9.307e-03   6.317 3.65e-10 ***
WoodDeckSF          9.309e-05  2.456e-05   3.791 0.000157 ***
ScreenPorch         2.249e-04  5.151e-05   4.365 1.37e-05 ***
SaleTypeNew         1.027e-01  1.488e-02   6.902 7.94e-12 ***
SaleConditionNormal 6.369e-02  1.010e-02   6.308 3.84e-10 ***
ConditionNorm       1.540e-01  3.482e-02   4.424 1.05e-05 ***
ExteriorBrkFace     5.539e-02  1.658e-02   3.340 0.000860 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1001 on 1318 degrees of freedom
Multiple R-squared:  0.9298,    Adjusted R-squared:  0.9278
F-statistic: 471.6 on 37 and 1318 DF,  p-value: < 2.2e-16
```

- Updated Stepwise BIC

  We re-run stepwise BIC and again reach the same model as with the one derived from ANOVA F-test.

- Updated Test for Normality

  To test the normality of our new best model, we calculate the standardized residual and then plot the qqplot of the standardized residuals



As we can see above in the Normal Q-Q Plot, the residuals looks pretty straight and fit the diagnosis line much better.

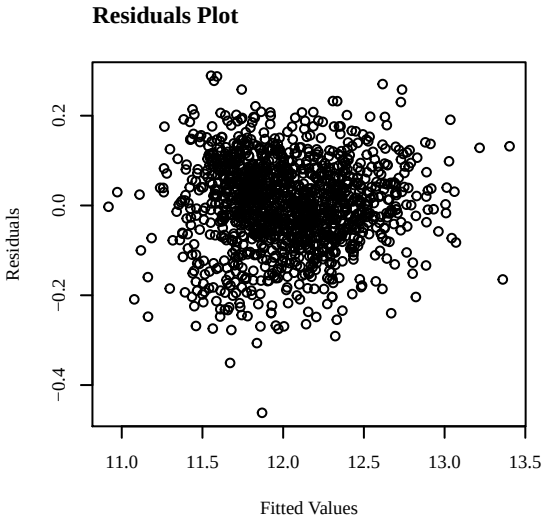
To further test the normality quantitatively, we iterate the Kolmogorov-Smirnov test 100 times and take the mean of 100 ks scores as our final ks score, which in our case is 0.299033. Since the mean ks score we get is very large, the normality assumptions for the data are satisfied as we fail to reject the null hypothesis that the residuals are normally distributed.

- Multi-colinearity Detection

```
      XMSSubClass50       XMSSubClass60       XMSSubClass70
           1.782950            3.226317            1.418108
        XMSZoningRL             XLotArea      XLandContourHLS
           1.602940            1.462080            1.109517
  XLotConfigCulDSac XNeighborhoodBrkSide XNeighborhoodCrawfor
           1.119556            1.248725            1.296190
XNeighborhoodNoRidge XNeighborhoodNridgHt XNeighborhoodSomerst
           1.283828            1.423919            1.548217
XNeighborhoodStoneBr        XYearRemodAdd         XExterQual.L
           1.131133            2.392249            2.930967
   XFoundationPConc          XBsmtFinSF1         XHeatingQC.L
           2.240583            2.035688            1.757137
    XCentralAirTRUE           XX1stFlrSF           XX2ndFlrSF
           1.272511            3.403784            1.943227
      XBsmtFullBath            XFullBath            XHalfBath
           1.822919            2.343232            1.939889
     XKitchenQual.L        XTotRmsAbvGrd        XFunctional.L
           2.635468            2.893422            1.086525
       XFireplaces    XGarageTypeAttchd          XGarageCars
           1.513969            1.544580            2.001846
     XPavedDrive.L          XWoodDeckSF         XScreenPorch
           1.262653            1.211822            1.080909
      XSaleTypeNew XSaleConditionNormal       XConditionNorm
           2.270180            1.910204            1.082452
   XExteriorBrkFace
           1.091753
```

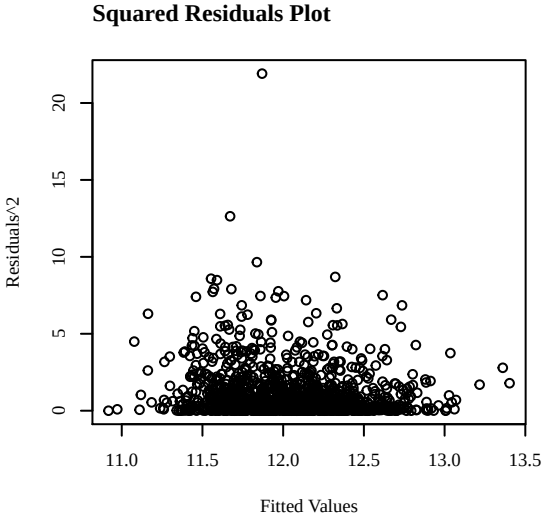
To detect any multi-colinearity issues in the model, we are calculating the variance inflation factors and using the rule of thumb as the criteria. None of the VIFs is going beyond 10. Thus, we believe there is no multi- collinearity among the variables in our model.

- test for Nonlinearity:
- plot residuals against fitted Ys



The residuals plot shows the residuals have no specific trend versus fitted values, indicating that the linear association may be appropriate.

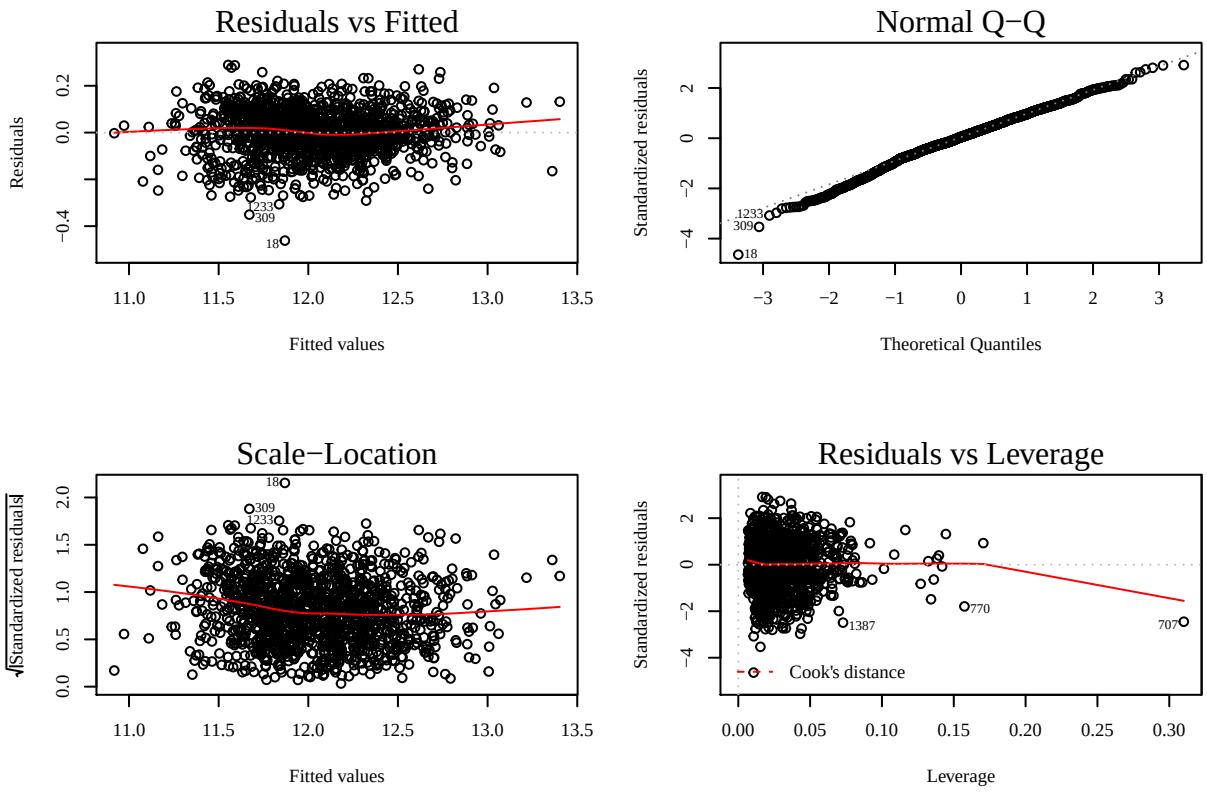
- test for constant variance:
  - plot residuals against fitted Ys
  - regress residuals on pairwise products of Xs, test stats: R^2
  - transformation $y = \sqrt{|y|}$ stablizes the variance



The squared residuals plot shows that the deviation of the residuals does not increase as the fitted values increase, indicating no nonconstancy of error variance.

2.1.5 Finalizing Model

Here is a summary plot of the best model



2.2 Model Application - Morty's Case

2.2.1 Selling Price Suggestion

Based on the model we built above, the 95% confident interval of prediction of the price Morty can sell his house for is *(128004.1, 190518.2)*, with the expected price of *156156163.7*.

If Morty tries hard enough and he gets a brilliant agent, he can sell his house for *190518.2* because that is the upper limit for the confidence interval of prediction, which means *190518.2* is not unlikely at the significant level of 95%.

Things can happen if Morty tries hard enough! :)

2.2.2 Price Enhancement Proposal

If Morty wants to sell his house for more, there are several things he can do to add value to this house based on the model: + The first thing Morty should work on is his kitchen. His kithen is so bad that no one wants to even takes a look. That's even probably the reason why they want to sell the house. + The second one is to upgrade the material on the exterior, which can increase the expected value of the house a lot. + Thirdly, Morty shoudl considier reconstruct the whole house. It may sound too costly, but the return is `huge`. The house will gain tons of value if the construction is new.

Morty's house is of high value already because the house already has some of the most significant features that ganruantee the price of the house, for example the house has the Central air conditioning and regarding the functionality, the house also has a very good performance.

So, let's see what happens if Morty works his ass off and upgrade the house like expected, the price would be:

Woo hoo! Now Morty is a very happy dude because after several changes, he can sell his house for *275909.6*! Morty has not seen that coming and now he truly believes the magic of data science and the power of experts.

You are welcome Morty.

3. Predictive Model

3.1 Data Cleaning

The data cleaning procedure for the predictive model should be different from the explanatory model. There are many quality features in the dataset and they are properly encoded as ordered factors in the explanatory model. However, this is quiet inefficient in the predictive point of view and brings a large loss of degree of freedom. Therefore we transform quality features to numerical scores (i.e. 'Excellent' to 5 and 'Poor' to 1), which is not a standard way to deal with categorical features but can lower MSPE.

3.2 Feature Engineering

In the explanatory model, Lasso is used to select variables. But the performence of Lasso to select variables is just fair and unstable. Sometimes it will drop variables that are valuable and increase MSPE. When n is large, we can drop variables in a more conservative way. In fact, for OLS we have

$$E(MSE) \approx \sigma^2(1 + p/n)$$

So even if we include a redundant variable it just increase MSE with σ^2/n. This is very small when n is big enough. In Lasso and Ridge, penalty on coefficients makes it even safer to keep most variables.

Therefore, instead of dropping variables aggressively with Lasso, we only drop a few categorical features that are highly uneven (most observations are in one category) and also not significant in t-test. Dropping these variables shows almost no influence on the cross validated MSPE but makes our model simpler and more stable.

Interaction terms can be added to the model to capture the non-linear and interactive patterns of the data. It is impossible to test all possible interaction terms, so we choose four most important features *Total Area*, *Zoning Type*, *Overall Quality* and *Neighborhood*, and consider all possible combinations of their 2-order and 3-order interactions. These interaction terms are transformed into hundreds of dummy features. Perhaps most of them are not significant but it is impossible to test all the dummy features one by one. Fortunately, Lasso can shrink the trivial variables automatically and cross validation results show that adding interaction terms into the model dramatically decrease the MSPE, which once again proves that redundant variables are not that important for predictive models.

Besides interaction terms, we create some features that may be useful in predicting housing price, e.g. average room area with total area divided by number of rooms. Some of them work and some of them don't. Features are only kept if they bring improvement to the CV result.

3.3 Model Fitting

We use glmnet to fit our model. We keep 30% of the total dataset as our validation data and the rest 70% as training data. Then we apply 10-fold cross validation within training data. We use the best lambda we get

to fit a new model and apply it to the validation set to get the validation score. As the data is relatively small, validation results are quiet unstable, so we run the above process 5 times and take the average as our final result.

The only parameter we need to tune is α. We apply a simple grid search procedure to test α from 0 to 1 with step 0.1.

3.4 Result

The result is listed below.

α	CV $MSPE$	std(CV)	Holdout $MSPE$	std(Holdout)	Holdout $\sqrt{MSPE}$ (Original Scale)
0	0.0192	0.0025	0.0152	0.0022	23527
0.2	0.0175	0.0029	0.0139	0.0016	21756
0.4	0.0206	0.0049	0.0139	0.0014	21539
0.6	0.0178	0.0046	0.0137	0.0020	21399
0.8	0.0170	0.0034	0.0135	0.0017	21374
1.0	0.0190	0.0048	0.0139	0.0015	21598

According to the table, we get lowest CV and validation results when $\alpha = 0.8$, although the MSPEs are stable for $\alpha > 0.2$. Ridge regression performs significantly poorer than Lasso because we include hundreds of interaction terms in the model. Ridge does not do any variable selection therefore the p is large and the variance of the model is high, while Lasso and elastic net shrink unimportant features to zero.

Note that although we have taken a 5-run-average, the result is still quiet unstable and may vary if we rerun the process. Outliers in the data dramatically influence the MSPE results in such a small dataset. Unfortunately we cannot drop outliers in holdout data otherwise the validation may not be correct (there are outliers in the real world).

We get dramatical improvement by introducing interactions within only 4 features. There are potential improvements by introducing other interactions, but we have already seen how the dimension blows up by including interactions in linear models. In this case, tree-base models, which inherently consider interactions and non-linearity of the data, or neural network, which introduces interactions by stacking multiple layers of linear models, may be better choices to predict housing prices for this dataset.