The Analysis of House Prices

1 Exploring data

The analysis described throughout the paper is using a subset of the full data set which was obtained by performing basic cleaning procedures to avoid issues with missing data and facilitate the analysis. We only selected those variables from the data set which did not have more than 10% of their values as missing. After removing these variables, the data set contained 75 features instead of the original 81. Then we selected those observations with complete entries for each of the 75 features selected. This reduced the number of observations from 1460 to 1338.

There are 36 continuous variables and 37 categorical variables in the data set. In linear regression, categorical variable is represented by a set of dummy variables that take value one when the observation is in that group and zero otherwise. If there are n groups for a categorical variable, there are $n-1$ dummy variables to avoid perfect multicollinearity since the sum of n dummy variables is one. The coefficient of a dummy variable is the difference of the mean of dependent variable of that group with the mean of dependent variable of the reference group. If that group contains few observations, then the coefficient estimate for the corresponding dummy variable is not significant since there are not enough samples and thus the variance of the coefficient estimate is high. Furthermore, if the categorical variable has multiple groups, it would cause perfect multicollinearity between dummy variables that represent other groups. Therefore, we encoded each factor variable as a set of dummy variables and selected dummy variables individually rather than selecting categorical variables in the feature selection process. For each categorical variable, we chose the group that has the most observations as the reference level of that categorical variable. We removed dummy variables whose group appears only in train set but not in test set since we need to apply the training model to the test set later on.

2 Assumptions of an linear model

We start with drawing a histogram of sale prices, which can be found in Figure 1.

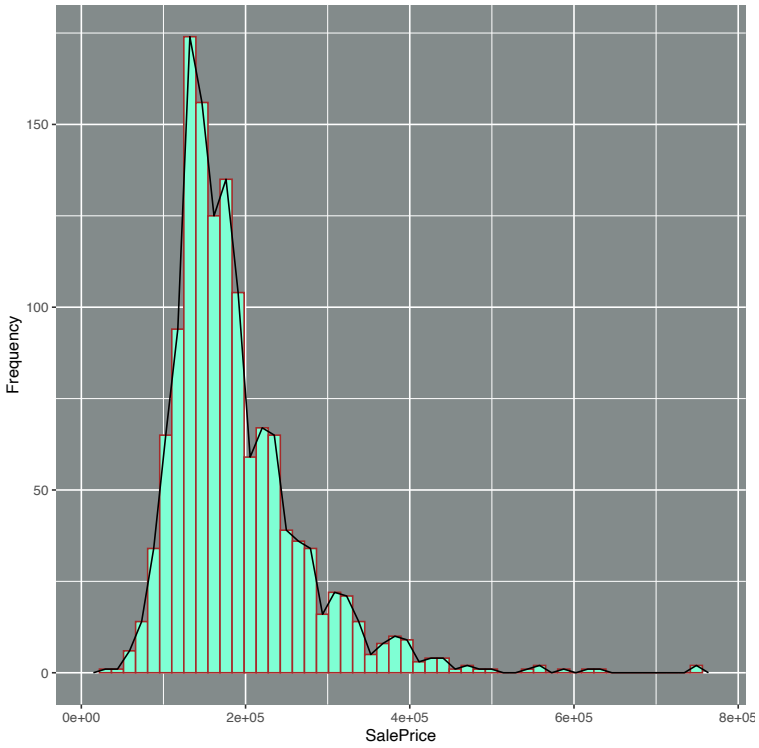

Figure 1: Histogram of the sale prices

Min.	1st Qu.	Median	Mean	3rd Qu.	Max
35311	135000	168500	186762	220000	755000

The distribution of Sale Price is skewed to the right as shown in Figure 1, which also supported by the summary statistics of Sale Price, as the mean is greater than the median. We use all of the variables available to run a basic regression model. Then we look at the residual plot and the normal QQ-plot. They can give use an idea of our normality assumptions. The deviance of the distribution from the normal distribution was further verified from the residual plot and normal QQ plot which can be found in Figures 2 and 3 respectively. The spread of residuals is small when fitted values are small and are not even around the X- axis when fitted values are large. This suggests the need for data transformation on the respondent variable. The upper and lower quantiles of the residuals are not aligned with the normal quantiles.

We used the box-cox procedure to a power transformation on the dependent variable. The optimal λ is the one that minimizes SSE, i.e., maximizes the log-likelihood when errors are assumed to be normally distributed. We run the box-cox procedure for different values of λ and then plot the log-likelihood for the model using that transformation on the data in Figure 4. The maximum value of the log-likelihood is reached when $\lambda = 0.14$. The 95% confidence interval for λ is $(0.10, 0.18)$.

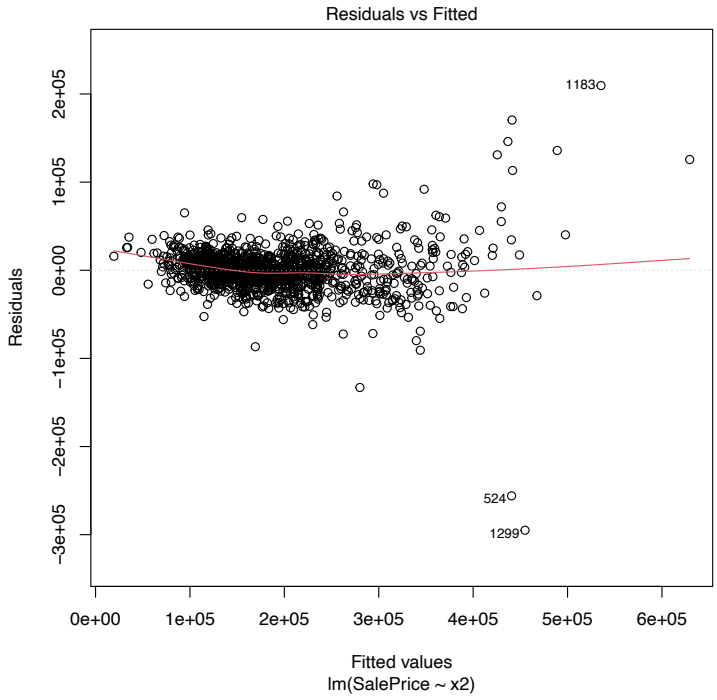

Figure 2: residual versus fitted values for the regression model

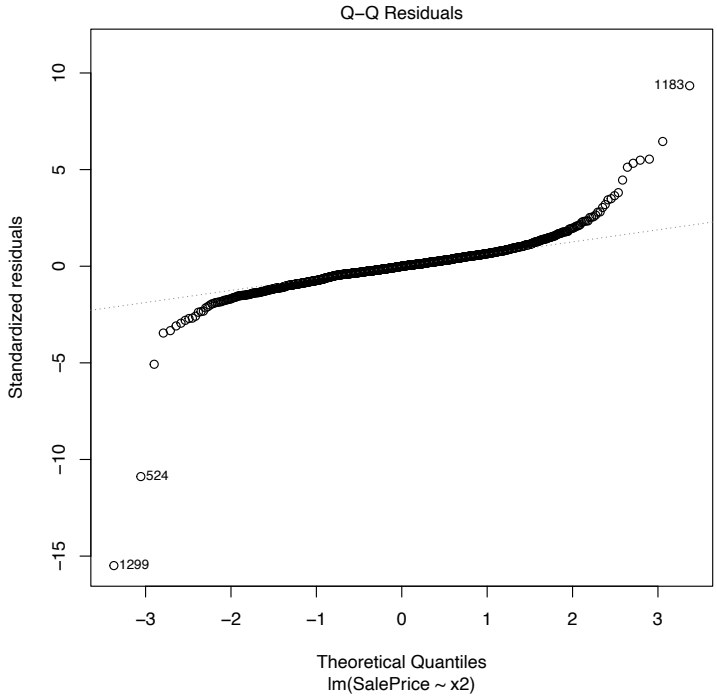

Figure 3: Quantiles of residuals versus their theoretical quantiles under normality

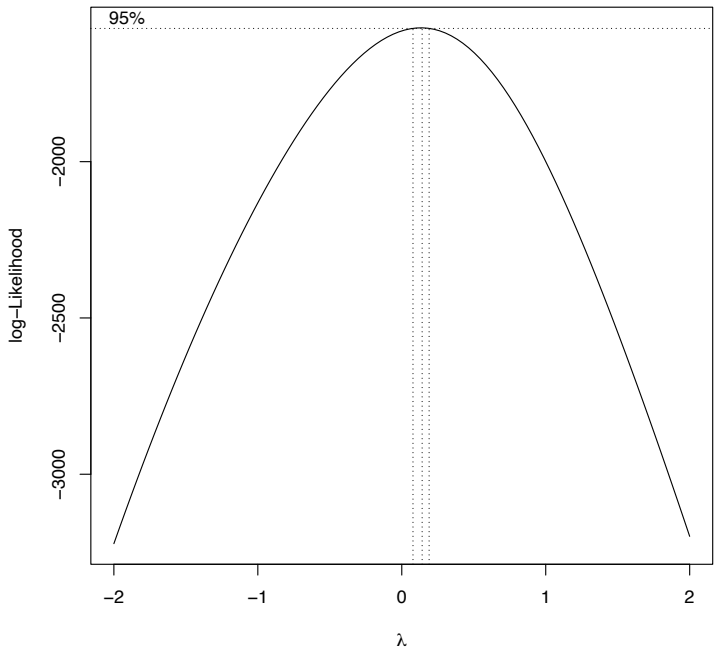

Figure 4: Log-likelihood plotted against different values of λ using the Box-Cox procedure

When $\lambda \neq 0$, the transformed $\tilde{Y} = \frac{Y^{\lambda}-1}{\lambda}$. Since we need to inversely transform $\tilde{Y}$ to get actual predicted $\hat{Y}$, we take $\lambda = 0.1$ since the power $\frac{1}{\lambda}$ in the inverse transform $Y = (\lambda\tilde{Y} + 1)^{1/\lambda}$ has to be integer to avoid complex numbers.

Inspecting the new studentised residual and QQ plot using the transformed data in Figure 5, there are still some points that are far away both in the residual and the QQ- plot, e.g. case number 826,633, 524. Since these points remain different after thee Box-Cox transformation, we would wish to explore these points more, In particular, we would like to know whether these points might be influential to the whole model.

3 Model Diagnostics and Influential Observations

In this section, We identify outlying Y observations through studentised deleted residuals and X observations through hat matrix leverage values Then we identify influential cases through DFFITS, Covariance ratio and Cook's distance.

We delete the ith case and fit a regression function on the remainder of the data points, Then we use it to estimate Y_i. Call that estimate $\hat{Y}_{i(i)}$. Then $d_i = Y_i - \hat{Y}_{i(i)}$ is called the deleted residual. This is also the PRESS statistic for the ith observation. The studentised deleted residual is given

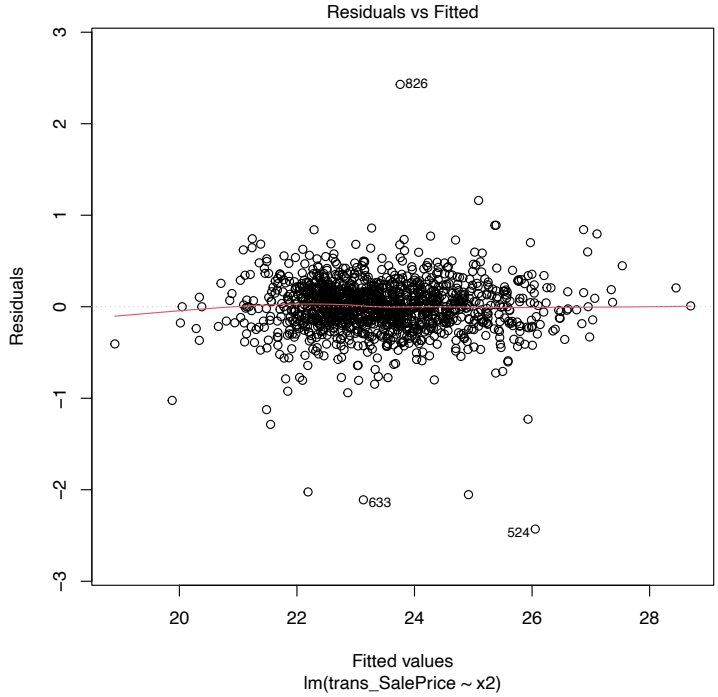

Figure 5: residual versus fitted values for the standard regression model using the transformed Y

by $\frac{d_i}{s(d_i)}$ which follows t-distribution of degree $n-p-1$. Here $s(d_i) = s_{(i)}\sqrt{1-h_{ii}}$ where $s_{(i)}$ is the sample deviation computed by deleting the ith observation. We can reject if our residual is greater than $t_{0.975,1114}$. We compute the studentised residuals. Looking at their summary statistics, there are 20 NA's since their hat value is one. Except those 20 cases, there are 59 outlying cases.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
-12.14341	-0.52486	0.04555	-0.00002	0.54665	12.14341	20

If the ith case is outlying in terms of X observations and therefore has a a large leverage value h_{ii}, it exercises substantial leverage in determining the fitted value of Y_i. Since h_{ii} is a function of the X values, it determines how important a particular observation X_i is in predicting Y_i. Leverage values are considered large if it is more than twice as large as the mean leverage value. So we check if $h_{ii} > 3\hat{h} = \frac{\sum_i^n h_{ii}}{n} = \frac{3p}{n}$, which in our case $n = 1338$ and $p = 224$ is 0.50. Checking the summary of leverages shows that most of the leverages to be below our given threshold. The exact cutoff point is around 5% for the threshold. There are 67 points that have leverage that exceeds the threshold.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.02264	0.07354	0.11312	0.16592	0.19122	1.00000

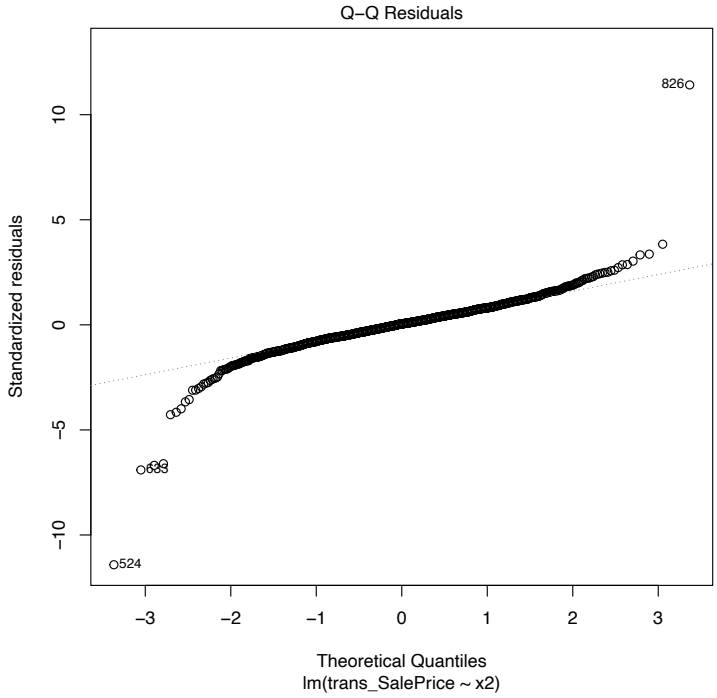

Figure 6: Quantiles of residuals versus their theoretical quantiles under normality using the transformed Y

90%	91%	92%	93%	94%	95%	96%	97%	98%	99%	100%
0.3368	0.3564	0.3800	0.4069	0.4451	0.4973	0.5504	0.5782	0.6890	1.0000	1.0000

After identifying cases that are outlying with respect to their Y values or X values, the next step is to see if an exclusion of a case can cause major changes in the fitted regression function, then such a case is identified as influential. We use three metrics DFFITS, Cook's distance and Covariance ratio.

DFFITS: A useful measure of the influence that the ith case has on the fitted value $\hat{Y}_i$ is given by $DFFITS_i = \frac{Y_i - \hat{Y}_{i(i)}}{MSE_{(i)}h_{ii}}$. A case is considered as influential if it absolute value of DFFITS exceeds $2\sqrt{p/n}$. In our case, that number is 0.82. We find an exact count of that number to be 45 cases influential and 1273 cases not influential.

Covariance Ratio: It is the measure of the impact of each observation on the variances and standard errors of the regression coefficients and their covariances. Values outside the interval $1 \pm 3p/n$ are considered highly influential. For our case, that range is $(0.50, 1.50)$. 140 out of 1338 cases fall outside that range.

Cook's distance: in contrast to the DFFITS which considers the influence of the ith cased on the fitted value Y_i for this case, Cook's distance measure $D_i = \frac{\sum_{i=1}^{n}(Y_j - Y_j(i))^2}{pMSE}$ considers the influence of the ith case on all n fitted values. We will call case i to be influential if $D_i > F_{0.50,p,n-p}$ where F represents the F-distribution. For our case, no point is influential.

There is a total of 178 cases that satisfy one or more outlying or influential criteria, which accounts for 178/1338 =13.3% of all cases being considered.

4 Method

After removing the outlying and influential cases, then we move to the next step of selecting features. Since dummy variable with less than 5 observations could easily cause perfect multicollinearity among predictors, we removed dummy variables of sample size that is less than five and category variable with only one group left before we apply feature selection algorithms. for the reason mentioned in section 1. There are backward and forward stepwise algorithms to do feature selection. The forward stepwise algorithm works by starting a null model without any predictors and adding one variable at a time to improve the AIC metric. The model selected by the forward stepwise algorithm has 86 variables. It turns out that many of which are nonsignificant. Therefore we choose the model selected by the backward stepwise algorithm. The backward stepwise algorithm starts with the full model with all features. The algorithm sequentially removes a predictor

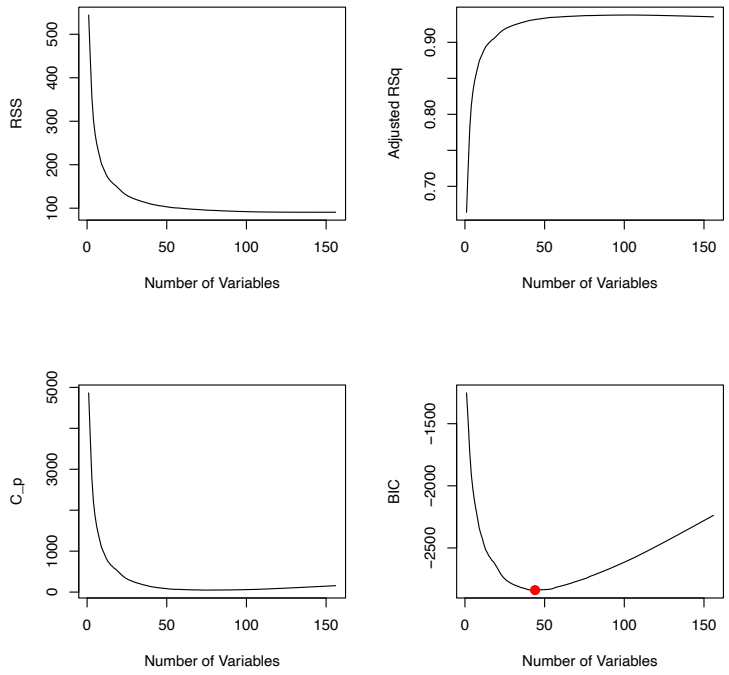

Figure 7: C_p, BIC, Adjusted R^2 and RSS of the best model selected by the backward stepwise selection algorithm given a fixed number of variables

that gives the smallest increase in RSS. Then we can select the best model based on BIC, adjusted R^2 or C_p given different sizes. C_p, BIC, Adjusted R^2 and RSS of all the models for each model size is shown in Figure 7. The RSS and is decreasing with the model size. The RSS, adjusted R^2, and C_p becomes stable when the number of variables gets 50. The BIC first decreases and then starts to increase. The minimum BIC is reached when the number of variables is 44, which is indicated by the red dot in the figure. The We choose the model with the smallest BIC as our final model.

```
Call:
lm(formula = formula1, data = house_df_f)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3622 -0.1722 -0.0018  0.1802  2.8018

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.689e+00  1.309e+00   3.583 0.000354 ***
MSSubClass          -2.215e-03  3.189e-04  -6.947 6.32e-12 ***
```

```
MSZoningRM           -1.360e-01  3.576e-02  -3.804 0.000150 ***
LotArea               1.205e-05  1.853e-06   6.503 1.19e-10 ***
LotShapeIR3          -4.710e-01  1.160e-01  -4.060 5.25e-05 ***
LotConfigCulDSac      1.331e-01  3.919e-02   3.398 0.000703 ***
NeighborhoodSomerst   2.613e-01  4.263e-02   6.128 1.23e-09 ***
NeighborhoodNridgHt   3.531e-01  5.006e-02   7.053 3.06e-12 ***
NeighborhoodEdwards  -2.974e-01  4.474e-02  -6.648 4.64e-11 ***
NeighborhoodCrawfor   4.145e-01  5.731e-02   7.233 8.78e-13 ***
NeighborhoodNoRidge   3.377e-01  5.801e-02   5.822 7.60e-09 ***
NeighborhoodStoneBr   4.773e-01  7.297e-02   6.541 9.26e-11 ***
NeighborhoodMeadowV  -4.080e-01  9.943e-02  -4.103 4.37e-05 ***
Condition1Feedr      -2.118e-01  4.593e-02  -4.611 4.47e-06 ***
Condition1Artery     -2.365e-01  5.954e-02  -3.973 7.56e-05 ***
BldgTypeTwnhs        -1.847e-01  6.334e-02  -2.917 0.003610 **
BldgTypeDuplex       -2.318e-01  7.453e-02  -3.110 0.001917 **
OverallQual           1.775e-01  1.356e-02  13.090  < 2e-16 ***
OverallCond           1.394e-01  1.111e-02  12.547  < 2e-16 ***
YearBuilt             7.556e-03  6.639e-04  11.380  < 2e-16 ***
Exterior1stHdBoard   -1.110e-01  2.702e-02  -4.107 4.31e-05 ***
Exterior1stAsbShng   -3.748e-01  1.019e-01  -3.680 0.000244 ***
Exterior2ndWdShng    -2.540e-01  6.494e-02  -3.912 9.72e-05 ***
Exterior2ndStucco    -6.274e-01  1.043e-01  -6.017 2.41e-09 ***
ExterCondGd          -9.036e-02  3.280e-02  -2.755 0.005972 **
FoundationCBlock     -9.749e-02  2.663e-02  -3.661 0.000263 ***
BsmtQualEx            1.502e-01  4.259e-02   3.527 0.000437 ***
BsmtCondFa           -2.027e-01  6.770e-02  -2.994 0.002818 **
BsmtExposureGd        2.097e-01  3.661e-02   5.727 1.31e-08 ***
BsmtFinType1Unf      -1.527e-01  2.541e-02  -6.011 2.49e-09 ***
HeatingQCTA          -7.070e-02  2.471e-02  -2.861 0.004302 **
CentralAirN          -2.544e-01  6.522e-02  -3.901 0.000102 ***
X1stFlrSF             9.346e-04  4.529e-05  20.636  < 2e-16 ***
X2ndFlrSF             8.100e-04  3.375e-05  24.001  < 2e-16 ***
BsmtFullBath          1.501e-01  2.257e-02   6.651 4.56e-11 ***
FullBath              1.007e-01  2.749e-02   3.664 0.000260 ***
KitchenQualEx         2.791e-01  4.566e-02   6.112 1.36e-09 ***
FunctionalMod        -3.558e-01  1.306e-01  -2.725 0.006536 **
Fireplaces            8.626e-02  1.799e-02   4.796 1.84e-06 ***
GarageCars            1.961e-01  2.129e-02   9.207  < 2e-16 ***
WoodDeckSF            2.737e-04  8.126e-05   3.369 0.000781 ***
```

```
ScreenPorch           1.082e-03  1.693e-04   6.388 2.47e-10 ***
PoolArea             -4.118e-03  4.529e-04  -9.092  < 2e-16 ***
SaleTypeNew           1.158e-01  3.798e-02   3.048 0.002357 **
SaleConditionAbnorml -3.223e-01  4.028e-02  -8.000 3.10e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3093 on 1115 degrees of freedom
Multiple R-squared:  0.9342,Adjusted R-squared:  0.9316
F-statistic: 359.9 on 44 and 1115 DF,  p-value: < 2.2e-16
```

All variables selected by the backward stepwise algorithm are significant. We can see that we have a very high R-squared, at 0.9342. And adjusted R-squared is 0.9316, almost the same as R-squared. This means that the model is good. Next we check the multicollinearity of predictors chosen by the backward stepwise algorithm. If a variable is highly correlated with other variables, then the variance of the coefficient estimate gets inflated. This can impact the significance power of the estimator and further inflate the variance of the estimator of test Y. In this model, all coefficient estimators have VIF less than 10, indicating that they are not highly correlated. Based on the t values, the most significant predictors are X2ndFlrSF and X1stFlrSF, followed by OverallQual, OverallCond and YearBuilt.

5 Lasso and Ridge Regression

We did lasso and ridge regression. We used cross-validation to select the penalty parameter λ. The λ used in our model was one-standard deviation higher the one that achieves the smallest cross validated MSE. For ridge regression, this number is 0.8188154. For lasso regression, this number is 0.1133966. Table 1 collects the coefficient estimates of three regression models. The majority of coefficient estimates from ridge and lasso regression models are smaller than those from the linear regression model. This is because lasso and ridge regression minimize the sum of L_0 and L_1 penalty term associated with coefficient estimates and SSE. Furthermore, only coefficients of 13 variables in the lasso regression model are nonzero, indicating that sparsity of lasso coefficients. The training and test MSE of three methods are shown in Table 2. The training MSE of lasso and ridge regression models are larger than the training MSE of linear regression, but ridge and lasso regression models have smaller test MSE than the linear regression model. We un-transform the predicted sale price back into the original scale and calculate the difference between the actual sale price and untransformed predicted sale price, which can be found in table 3. Even though MSE of models using transformed data is small, MSE calculated from the difference between un-transformed predicted values and actual sale prices is large.

6 Variable Importance by Bagging Procedure

For comparison, we also tried other method to select features. Regression trees are a powering tool to measure the importance of a predictor through its reduction on SSR. Bootstrap aggregating or bagging is a method of averaging model performances. We partition the training set of size n into m new training sets. Then we run m many regression trees and then make the final model by averaging. We will use this for variable selection, by a leave-out-out procedure,

	Standard	Lasso	Ridge
(Intercept)	4.689e+00	1.552e+01	1.323e+01
MSSubClass	-2.215e-03	0.000e+00	-1.010e-03
MSZoningRM	-1.360e-01	-1.918e-01	-2.101e-01
LotArea	1.205e-05	6.527e-06	1.195e-05
LotShapeIR3	-4.710e-01	0.000e+00	-2.163e-01
LotConfigCulDSac	1.331e-01	0.000e+00	1.364e-01
NeighborhoodSomerst	2.613e-01	0.000e+00	1.556e-01
NeighborhoodNridgHt	3.531e-01	0.000e+00	2.743e-01
NeighborhoodEdwards	-2.974e-01	0.000e+00	-2.458e-01
NeighborhoodCrawfor	4.145e-01	0.000e+00	2.761e-01
NeighborhoodNoRidge	3.377e-01	0.000e+00	4.415e-01
NeighborhoodStoneBr	4.773e-01	0.000e+00	3.266e-01
NeighborhoodMeadowV	-4.080e-01	0.000e+00	-3.565e-01
Condition1Feedr	-2.118e-01	0.000e+00	-1.363e-01
Condition1Artery	-2.365e-01	0.000e+00	-1.662e-01
BldgTypeTwnhs	-1.847e-01	0.000e+00	-1.787e-01
BldgTypeDuplex	-2.318e-01	0.000e+00	-1.497e-01
OverallQual	1.775e-01	3.690e-01	1.496e-01
OverallCond	1.394e-01	0.000e+00	5.081e-02
YearBuilt	7.556e-03	2.029e-03	3.781e-03
Exterior1stHdBoard	-1.110e-01	0.000e+00	-6.790e-02
Exterior1stAsbShng	-3.748e-01	0.000e+00	-2.769e-01
Exterior2ndWdShng	-2.540e-01	0.000e+00	-2.220e-01
Exterior2ndStucco	-6.274e-01	0.000e+00	-3.287e-01
ExterCondGd	-9.036e-02	0.000e+00	-1.473e-02
FoundationCBlock	-9.749e-02	0.000e+00	-1.184e-01
BsmtQualEx	1.502e-01	1.067e-01	2.561e-01
BsmtCondFa	-2.027e-01	0.000e+00	-2.344e-01
BsmtExposureGd	2.097e-01	0.000e+00	1.891e-01
BsmtFinType1Unf	-1.527e-01	0.000e+00	-9.503e-02
HeatingQCTA	-7.070e-02	-1.477e-02	-1.333e-01
CentralAirN	-2.544e-01	0.000e+00	-3.117e-01
X1stFlrSF	9.346e-04	5.840e-04	4.162e-04
X2ndFlrSF	8.100e-04	2.552e-04	3.257e-04
BsmtFullBath	1.501e-01	3.499e-02	1.120e-01
FullBath	1.007e-01	1.270e-01	2.346e-01
KitchenQualEx	2.791e-01	1.511e-02	3.188e-01
FunctionalMod	-3.558e-01	0.000e+00	-1.899e-01
Fireplaces	8.626e-02	7.867e-02	1.445e-01
GarageCars	1.961e-01	2.898e-01	2.163e-01
WoodDeckSF	2.737e-04	0.000e+00	4.164e-04
ScreenPorch	1.082e-03	0.000e+00	8.345e-04
PoolArea	-4.118e-03	0.000e+00	-1.781e-03
SaleTypeNew	1.158e-01	0.000e+00	1.524e-01
SaleConditionAbnorml	-3.223e-01	0.000e+00	-2.401e-01

Table 1: coefficient estimates of three regression models

	Standard	Lasso	Ridge
training MSE	0.3032085	0.5065056	0.3778004
test MSE	1.1852821	0.9350668	1.0081971

Table 2: Training and test MSE of three regression models

	Standard	Lasso	Ridge
training MSE	18277.36	33141.77	24366.06
test MSE	74027.96	54416.00	60377.48

Table 3: Training and test MSE of un-transformed sale prices

We leave each variable one and them take the ratio of the decreased SSR of the bagged trees to the full model SSR to find reduction in the SSR's from the full model. For the resulting regression tree, the ratio of SSR is given as a table in Table 4. There are 54 variables that are used to partition X space into rectangular regions in the full model. In regression tree model, The most important predictor is OverallQual, followed by GarageCars, X1stFlrSF and GarageArea. GarageArea was not selected by the backward stepwise algorithm. We plotted added-variable plots for eight most important variables selected by bagging for illustration purpose. As seen from Figure 8, there is partial correlation between SalePrice and GarageArea is weak, which is further supported by the p value of t test of the model. Since the algorithm only selects variables based on the strength of its linear relation with the target variable, GarageArea was not selected but has a significant nonlinear relationship with SalePrice in tree models.

```
Call:
lm(formula = newformula1, data = house_df_f)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0452 -0.2418  0.0282  0.3134  1.4665

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.852e+01  7.601e-02 243.605  < 2e-16 ***
OverallQual 4.112e-01  1.663e-02  24.721  < 2e-16 ***
GarageCars  3.579e-01  4.514e-02   7.928 5.23e-15 ***
X1stFlrSF   1.028e-03  5.959e-05  17.256  < 2e-16 ***
GarageArea  2.378e-04  1.553e-04   1.531    0.126
LotArea     1.694e-05  2.683e-06   6.315 3.84e-10 ***
X2ndFlrSF   6.333e-04  4.323e-05  14.648  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5076 on 1153 degrees of freedom
```

	variable	importance
1	OverallQual	1.68
2	GarageCars	1.18
3	X1stFlrSF	1.15
4	GarageArea	1.13
5	LotArea	0.93
6	BsmtFinSF1	0.87
7	YearBuilt	0.76
8	X2ndFlrSF	0.73
9	FullBath	0.69
10	MasVnrArea	0.56
11	BsmtQualEx	0.46
12	TotRmsAbvGrd	0.45
13	WoodDeckSF	0.43
14	BsmtQualGd	0.41
15	ExterQualGd	0.38
16	KitchenQualGd	0.38
17	BsmtFullBath	0.36
18	NeighborhoodClearCr	0.36
19	KitchenQualEx	0.36
20	BsmtUnfSF	0.34
21	LotConfigCulDSac	0.34
22	GarageYrBlt	0.32
23	ExterQualEx	0.29
24	GarageTypeDetchd	0.29
25	YearRemodAdd	0.28
26	Fireplaces	0.27
27	ExterCondGd	0.27
28	YrSold	0.27
29	BsmtExposureGd	0.27
30	HouseStyle2Story	0.26
31	NeighborhoodNoRidge	0.25
32	ScreenPorch	0.25
33	MoSold	0.25
34	BedroomAbvGr	0.25
35	SaleTypeNew	0.24
36	MSSubClass	0.24
37	Exterior1stWd Sdng	0.24
38	HalfBath	0.23
39	NeighborhoodCollgCr	0.22
40	OpenPorchSF	0.22
41	BsmtFinType1Unf	0.21
42	RoofStyleFlat	0.21
43	MSZoningRM	0.21
44	LandSlopeMod	0.21
45	OverallCond	0.19
46	LandContourLow	0.19
47	LotShapeIR1	0.18
48	LotShapeIR2	0.17
49	RoofStyleHip	0.17
50	LotConfigCorner	0.17
51	NeighborhoodNridgHt	0.15
52	CentralAirN	0.15

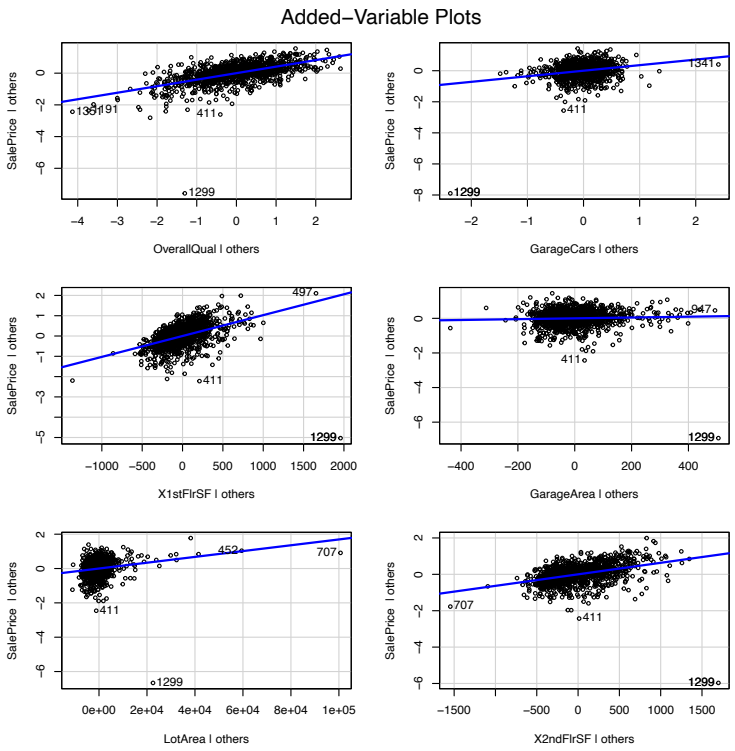

Figure 8: Added variable plots for eight most important variables selected by bagging

```
Multiple R-squared:  0.8168,Adjusted R-squared:  0.8158
F-statistic: 856.6 on 6 and 1153 DF,  p-value: < 2.2e-16
```

7 Discussion and Improvement

When we did the transformation on the dependent variable Y, the optimal λ was chosen for the full model. After the feature selection process, we can apply box-cox procedure to the model with selected variables to find an appropriate λ since the optimal λ that minimizes SSE for the new model might be different from the optimal one for the full model. It may help improve the training and test MSE calculated from residuals between actual sale prices and inversely transformed predicted sale prices.

Secondly, before feature selection, we removed the outlying and influential cases for the full linear regression model. The percentage of outlying and influential cases is 13.30%. It indicates that linear model may not be the best fit to capture the relation between the response variables and the variables within the scope to be considered. We may have missed information or patterns in those cases that may well capture similar patterns in the test data set. Therefore, we would consider different types of models.

In this work, we apply the whole training set to do feature selection. A more robust approach is to split the training set into k-fold, and get importance of features using cross validation.

In summary, we developed a model that was not only able to predict house prices but is also capable of explaining the nature of the relationship between specific variables and the response variable.