Math 151: Fundamentals of Data Science
Dr. Cristina Tortora

Regression: Predicting Housing Prices

Shengsheng Huo, Monica Orme, Robert Yav

May 9, 2022

Introduction

Currently, housing prices are getting more rigid, especially when we compare the housing prices in metropolitan areas such as the San Francisco Bay Area with houses in the other states. We specifically investigated housing prices of residential homes in Ames, Iowa in a house prices dataset named from Kaggle. In this project, we aim to build a model that accurately predicts sales prices of homes and investigate which factors will affect the price of houses in the United States. This dataset has 81 variables (columns) and 1460 observations (rows).

We will apply the multiple linear regression, kernel regression, regression tree, and k-means models to determine the best model that predicts the housing sales prices and accomplish the outlined goals. Before applying any of these models, we need to execute the data preparation.

Data Description

[1]In our project, we have one response. The response is the SalePrice. Most of the models in this projects considered the 25 variables below (not including the predictor response):

TotalBsmtSF - Is the total basement square footage.
GarageArea - Is the area of the garage in the house.
MoSold - Is the month that the house was sold (MM).
YrSold - Is the year that the house was sold (YYYY).
GrLivArea - Is the above grade (ground) living area in square feet.
BedroomAbvGr - Is the above grade (ground) bedroom area in square feet.
KitchenAbvGr - Is the above grade (ground) kitchen area in square feet.
TotRmsAbvGrd - Is the total amount of rooms that are above grade (does not include bathrooms).
GarageCars - Is the size of the garage in car capacity.
LotArea - Is the lot size area in square feet.
LotFrontage - Is the linear feet of the street connected to property.
YearBuilt - Is the original construction date.
OverallCond - Is the overall condition rating of the house.
MasVnrArea - Is the masonry veneer area in square feet.
BsmtFinSF2 - Is the type 2 finished basement area in square feet.
BsmtUnfSF - Is the unfinished basement area in square feet.
X1stFlrSF - Is the first floor area in square feet.
X2ndFlrSF - Is the second floor area in square feet.
WoodDeckSF - Is the wood deck area in square feet.
OpenPorchSF - Is the open porch area in square feet.
BsmtFullBath - Is the number of full bathrooms in the basement.
Fireplaces - Is the number of fireplaces.
FullBath - Is the number of full bathrooms that are above grade.
HalfBath - Is the number of half bathrooms that are above grade.

[1] The complete data description is attached to the project submission as a csv file and it describes the predictors that have been used in the dataset.

Data Preparation

Before the data could be analyzed, cleaning the data was necessary to prepare it for data analysis. The steps that we took to do this were to investigate whether the columns have appropriate variables, the rows that contained many missing values, and the distribution of the variables. After these steps were completed, the data were imputed.

We initially cleaned the data by verifying the variable types of the columns. The categorical variables had the character variable type, so these were all changed to the factor variable type. This change ensures that these are properly considered categorical in the dataset, data imputation, and data analysis. Next, we further cleaned the data by removing rows with at least 20% missing values since it would be difficult to maintain the data's distribution during imputation if there are observations with many missing values. Columns with at least 80% missing values were removed since most homes did not have the features mentioned for those columns and we are trying to predict a majority of homes. The removed columns include Fence, Alley, Miscfeatures, PoolQC. After this point, we looked at the distribution of each continuous variable using boxplots and we removed variables that had strongly skewed data or many outliers. These removed variables include BsmtFinSF2, ManvnrArea, and LotArea. We removed the extreme sales prices since we aim to predict housing prices on the average home. Finally, the data were imputed using multiple imputation. This method interpolates the missing data while maintaining the overall distribution of the data as much as possible.

Analysis Approach

The first approach is to create a regression model by using the forward selection method. We applied these variables and included some categorical variables to a multiple linear regression model. After this section, we only use the numerical variables to do backward selection to predict the multiple regression model. The second approach is to create decision tree algorithms to optimize variable selection by alternating the number of nodes of the tree. The third approach is to create kernel regression by changing the bandwidth of the model. The final approach is to create random forest regression by changing the number of mtry for prediction. After creating all these models, we try to compare all of its MSE and see which one is the better model to predict the sale price of the house.

Methodology

We have fit four different models and compared them by MSE (Mean Squared Error) to see which model would work best for predicting housing prices. We subset the cleaned data into a training and test dataset, where the training subset was a random sample of 70% of the data while the test was the remaining 30%.

The first model we attempted was a multiple linear regression model. We performed a stepwise forward variable selection to select meaningful parameters in the multiple linear regression model and obtained an MSE of 373589470. We also created a decision tree and let the algorithm optimize the variable selection. The MSE obtained was 487942403891, much higher than the multiple linear regression MSE. There was also pruning based on cross-validation, which actually increased the MSE of the model drastically. Next, we compared kernel regression. The model's bandwidth was tuned based on minimizing the MSE, which resulted in a final MSE of 3479005530. This was also much higher than the multiple linear regression MSE. Finally, we compared the random forest regression model. The model's number of predictors to consider for splitting were tuned using the MSE and the test MSE was 486558431 while the training MSE was 93262941. The multiple linear regression model had the lowest MSE and the random forest regression model had the second lowest MSE. Hence, we will primarily consider these models in the data analysis.

We chose to do the multiple linear regression on the continuous data to reduce the amount of interaction that would need to be calculated and simplify the model interpretation. We ran backward variable selection this time, which removed 8 insignificant variables. Backward Variable Selection is a technique where significant predictors are chosen using a certain threshold α. We begin with the full model using all predictors and drop predictors for each iteration based on the highest p-value. This algorithm stops when all predictors are less than or equal to the threshold α. We used α = 0.05 as the threshold to minimize the Type I error. We ended variable selection with the following model where y denotes the house sales price:

[2] $$y = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8+\beta_9x_9+\beta_{10}x_{10}+\beta_{11}x_{11}+\beta_{12}x_{12}+\beta_{13}x_{13}$$

Before the model can hold any significance in interpretation, we must first check assumptions. The assumptions of the multiple linear regression model are as follows (James, 2013):

1. The errors are normally distributed
2. The errors have constant variance
3. The errors are uncorrelated
4. There is linearity between the response and the predictors.

We will begin the assumption verification for the model below:

Assumption 1 - Normality: The Q-Q plot looks relatively normal, therefore the model checks the assumption of normality.

[2] - See Figure 7 under the Results section for a description of what the variables and parameters refer to

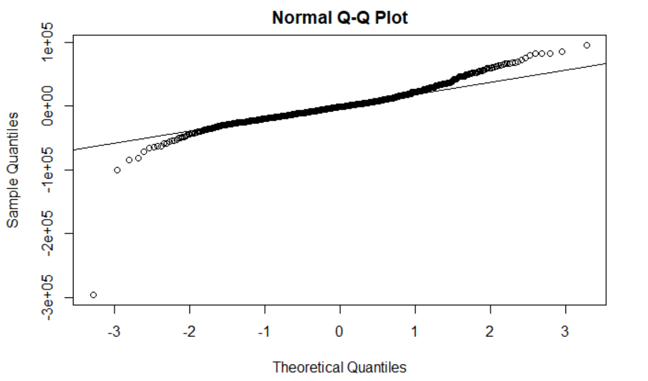


Figure 1: Normal QQ-Plot

Assumption 2 - Constant Variance: After plotting the fitted values against the residuals, there is obviously a shape taking place, therefore the constant variance assumption has been violated.

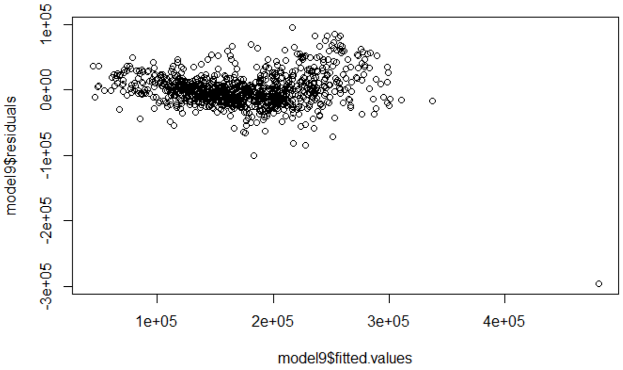


Figure 2: Fitted Values against the Residuals

Assumption 3 - Uncorrelated Errors: The residual plot also looks a bit suspicious, since much of the residuals seem to be crowding a center line instead of being randomly scattered. The uncorrelated error assumption is also violated.

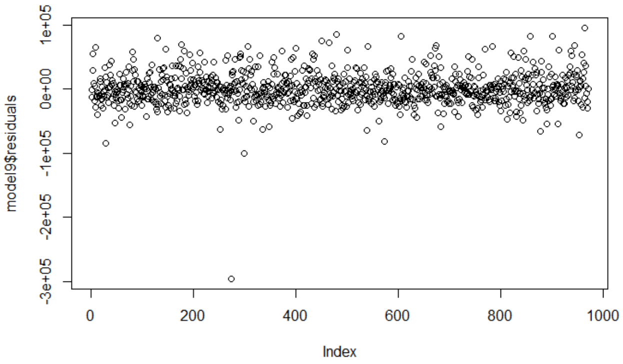


Figure 3: Residual Plot

Assumption 4 - Linearity: We performed an assumption of linearity check by plotting the response against each of the predictors and found that the model also violates the linearity assumption.

Since the constant variance assumption has been violated, we applied a log transformation to the response variable. In doing so, the variable "BedroomAbvGr" seemed to have increased in p-value to the point that it no longer became significant (Figure 8). The resulting transformed model was:

[3] $\log(y)=\beta0+\beta1x1+\beta2x2+\beta3x3+\beta4x4+\beta5x5+\beta6x6+\beta7x7+\beta8x8+\beta9x9+\beta10x10+\beta11x11+\beta12x12+\beta13x13$

In order to verify the model's validity, we must also verify the assumptions that were stated at the beginning of this section.

Assumption 1 - Normality: The Q-Q plot of the transformed model also looks straight and normal, satisfying the normality assumption.

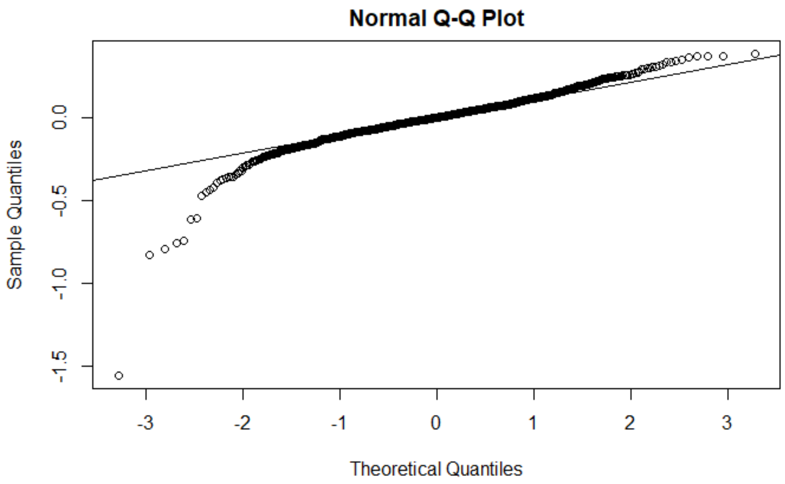


Figure 4: Transformed Model Normal QQ-Plot

Assumption 2 - Constant Variance: Plotting fitted values against the residuals show no significant difference between the transformed model and the old model, still violating the constant variance assumption.

[3] - See Figure 7 under the Results section for a description of what the variables and parameters refer to

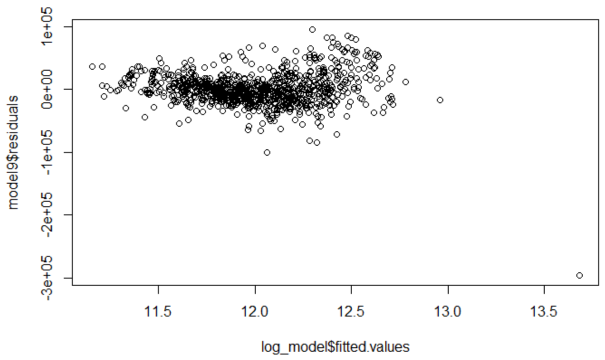


Figure 5: Transformed Model Fitted Values against the Residuals

Assumption 3 - Uncorrelated Errors: The plotted residuals don't show any significant improvement either, still violating the uncorrelated variance assumption.

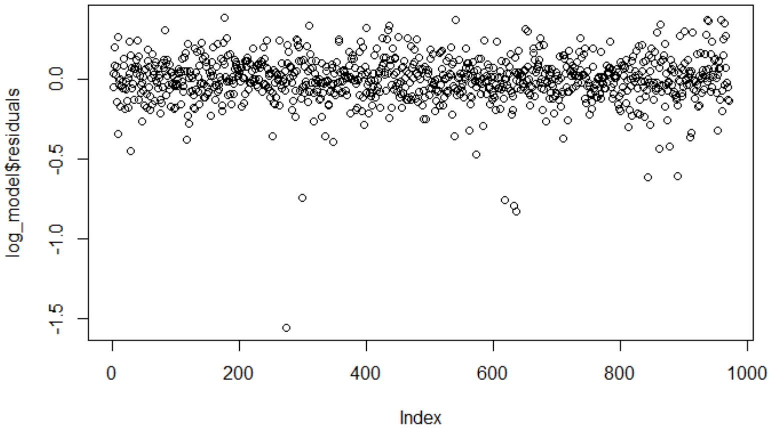


Figure 6: Transformed Model Residual Plot

Assumption 4 - Linearity: We also performed an assumption of linearity check by plotting the residuals against the predictors for each predictor and found that the model also violates the linearity assumption.

Since most of the assumptions were still violated, we performed k-means clustering on the data in order to create different regression models for each cluster, so that we can predict house prices for homogeneous groups. To get started, we computed the C(G) ratios for different amounts of clusters. The C(G) ratio is a measure of cluster strength and we aim to maximize it by choosing the highest ratio. The average silhouette width across potential clusters was also considered to determine how well the clusters were separated (Everitt, 2012). We decided on 5 clusters since the average silhouette width of 0.544 was slightly larger compared to that of 6 clusters (average silhouette width was 0.543). We then proceeded to perform multiple linear regression on each cluster. However, the R^2 of each model turned out to be less than 0.3, therefore the accuracy of each cluster model is questionable. Although the assumptions aren't completely met, the transformed-log multiple linear regression model was one of the closest models we performed that could predict the house prices.

Since most of the logarithmic transformed model's assumptions were violated, we considered the random forest regression tree model. This model builds a certain amount of decision trees on bootstrapped samples from the training dataset and a random sample of m predictors from the entire amount of p predictors is taken whenever a tree split is considered. This split only considers one of the m predictors. The value of m is often tuned before it is incorporated in the random forest regression tree (Everitt, 2012). This has been tuned using the test MSE and we selected m value that yielded the lowest MSE without overfitting, which was 10 in this case. Once this was completed, we proceeded with the analysis of the model in the results section. The random forest regression tree model that was developed utilized the housing sale price as the response and the remaining variables as predictors including categorical variables.

Results

Model 1: Logarithmic Multiple Linear Regression Model

According to the three models we applied here, we can see that

$\log(y)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8+\beta_9x_9+\beta_{10}x_{10}+\beta_{11}x_{11}+\beta_{12}x_{12}+\beta_{13}x_{13}$

The final model that we chose was the transformed log multiple linear regression model. We found that it violated the constant variance assumption, linearity assumption, and uncorrelated variance assumption. Although this model was chosen, there can't be any meaningful analysis done due to the violations.

Strictly within the model, the Log sales price can be predicted using the parameters (beta coefficients) from this model in the figure below:

Variable x_i	Variable Name	Beta Coefficient β_i	Interpretation
Intercept	Intercept	1.903e+00	For a theoretical house with a value 0 in all other variables, the average log house price would be $1.903 holding all other predictors fixed.
x1	TotalBsmtSF: total basement square footage	3.365e+01	For each point increase in TotalBsmtSF, the average log house price would increase by $3.365e+01 holding all other predictors fixed.

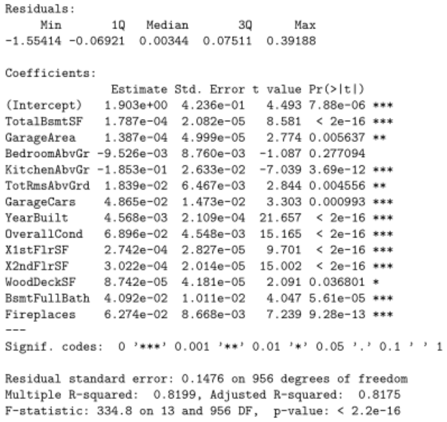
x2	GarageArea	3.311e+01	For each point increase in GarageArea, the average log house price would increase by $3.311e+01 holding all other predictors fixed.
x3	BedroomAbvGr	-6.266e+03	For each point increase in BedroomAbvGr, the average log house price would change by -$6.266e+03holding all other predictors fixed.
x4	KitchenAbvGr	-3.653e+04	For each point increase in KitchenAbvGr, the average log house price would change by -$3.653e+04 holding all other predictors fixed.
x5	TotRmsAbcGr	3.943e+03	For each point increase in TotRmsAbcGr, the average log house price would increase by $3.943e+03 holding all other predictors fixed.
x6	GarageCars	5.226e+03	For each point increase in GarageCars, the average log house price would increase by $5.226e+03 holding all other predictors fixed.
x7	YearBuilt: the year that the home was built	6.477e+02	For each point increase in YearBuilt, the log house price would increase by $6.477e+02 holding all other predictors fixed.
x8	OverallCond: the overall condition score of the home	8.759e+03	For each point increase in OverallCond, the average log house price would increase by $8.759e+03 holding all other predictors fixed.
x9	X1stFlrSF: the square footage of the first floor	5.014e+01	For each point increase in X1stFlrSF, the

			average log house price would increase by $5.014e+01 holding all other predictors fixed.
x10	X2ndFlrSF: the square footage of the second floor	5.777e+01	For each point increase in $X2ndFlrSF, the average log house price would increase by $5.777e+01 holding all other predictors fixed.
x11	WoodDeckSF: wood deck square footage	2.086e+01	For each point increase in WoodDeckSF, the average log house price would increase by $2.086e+01 holding all other predictors fixed.
x12	BsmtFullBath: number of bathrooms in the basement	7.093e+03	For each point increase in BsmtFullBath, the average log house price would increase by $7.093e+03 holding all other predictors fixed.
x13	Fireplaces: the number of fireplaces	8.819e+03	For each point increase in Fireplaces, the average log house price would increase by $8.819e+03 holding all other predictors fixed.

Figure 7: Transformed Model Parameter Interpretation

We can also interpret the adjusted R^2 for the transformed model from figure 8. The adjusted R^2 is a measure of the model's fit and it penalizes multiple regression models for each additional predictor. The adjusted R^2 for the model was 0.8175, which means that 81.75% of the variation in the log(Sales Price) is explained by the model, which indicates that the model does a good job of this since more than half of the variation is explained.

Another way to assess the model's goodness of fit is to look at the p-values of the beta coefficients using the threshold $\alpha = 0.05$. Beta coefficients with p-values equal to or less than the threshold indicate that there is sufficient evidence that the changes in the corresponding predictor are associated with changes in the log(sales price). Based on figure 8, we can note that the p-values for the beta coefficients for the Intercept, TotalBsmtSF, GarageArea, KitchenAbvGr, TotRmsAbcGr, GarageCars, YearBuilt, OverallCond, X1stFlrSF, X2ndFlrSF, WoodDeckSF, BsmtFullBath, and Fireplaces all have sufficient evidence supporting that the changes in the predictors are associated with changes in the log(sales price) since the p-values are below the threshold. However, for the p-value of BedroomAbvGr's beta coefficient, the p-value is above the threshold. This means that there is not sufficient evidence to conclude that changes in this predictor are associated with changes in the response.



```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.55414 -0.06921  0.00344  0.07511  0.39188 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.903e+00  4.236e-01   4.493 7.88e-06 ***
TotalBsmtSF   1.787e-04  2.082e-05   8.581  < 2e-16 ***
GarageArea    1.387e-04  4.999e-05   2.774 0.005637 ** 
BedroomAbvGr -9.526e-03  8.760e-03  -1.087 0.277094    
KitchenAbvGr -1.853e-01  2.633e-02  -7.039 3.69e-12 ***
TotRmsAbvGrd  1.839e-02  6.467e-03   2.844 0.004556 ** 
GarageCars    4.865e-02  1.473e-02   3.303 0.000993 ***
YearBuilt     4.568e-03  2.109e-04  21.657  < 2e-16 ***
OverallCond   6.896e-02  4.548e-03  15.165  < 2e-16 ***
X1stFlrSF     2.742e-04  2.827e-05   9.701  < 2e-16 ***
X2ndFlrSF     3.022e-04  2.014e-05  15.002  < 2e-16 ***
WoodDeckSF    8.742e-05  4.181e-05   2.091 0.036801 *  
BsmtFullBath  4.092e-02  1.011e-02   4.047 5.61e-05 ***
Fireplaces    6.274e-02  8.668e-03   7.239 9.28e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1476 on 956 degrees of freedom
Multiple R-squared:  0.8199,	Adjusted R-squared:  0.8175 
F-statistic: 334.8 on 13 and 956 DF,  p-value: < 2.2e-16
```

Figure 8: Transformed Model Output

Model 2: Random Forest Regression Tree

We developed a random forest regression model with the housing sales price as the response and the remaining variables as predictors. To assess this model, we considered the MSE on both the training and test samples, the percentage of variance explained by the model, variance importance plots, and a plot of the predicted sales price against the test housing prices. The training MSE was 93262941 and the test MSE was 486558431. The MSE for both are larger since many predictors and observations are contained in this model, which increases the variability. Furthermore, 85.07% of the variance was explained by the model, which shows the model is sufficient and describes the model's variability well.

To further assess the model's viability, we plotted the predicted sales price against the test housing prices to determine how well the model predicted the housing sales prices.

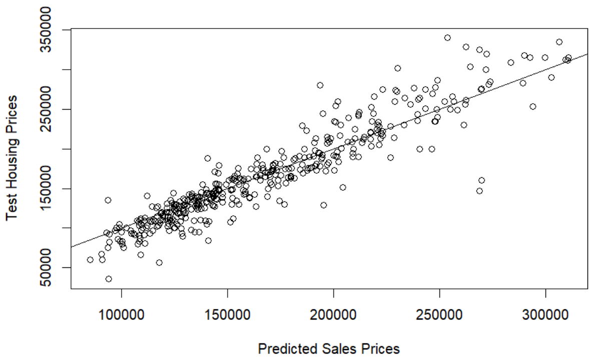


Figure 9: Predicted Housing Sales Prices vs. Test Housing Sales Prices

Based on the plot in figure 9, we can observe that there is a linear trend between the predicted housing sales price against the test housing prices. This means that many of the predicted response values were close to the actual housing sales prices. However, it is important to note that as the test housing prices increase, the predicted response values deviated more from the actual response values. Overall, this model viably predicts the response values since these tend to be close to the actual response values.

We also created variance importance plots to assess which variables were the most important in the model based on the percentage increase in the MSE (%incMSE). The variables with the highest percentage contributed the most to the model and these were GrLivArea and TotalBsmtSF since these had the highest percentage increase in the MSE. GrLivArea led to approximately a 40% percent increase in the MSE, while TotalBsmtSF led to approximately a 30% percent increase in the MSE.

Conclusion

The model that best predicted the housing sales prices was the random forest regression model since it explained most of the variance in the model and had one of the lower mean square errors out of most of the models that we did. The next best model was the logarithmic transformed multiple linear regression since the adjusted R^2 was high and the most of the predictors significantly affected the change in the logarithm of the housing sales prices. While the multiple linear regression model on the log sales price was done on the data, it violated the constant variance assumption, linearity assumption, and uncorrelated variance assumption, which vastly reduces the model's viability in predicting the response. Therefore, even though we developed a linear model that predicts housing sale prices in the United States, it cannot be used to accurately predict housing prices for any future values.

Discussion and Recommendation

Interesting discoveries were made during the data analysis process including that most of the variation in sales prices can be explained by some of the models. The random forest regression model was concluded as the best model for predicting prices due to the lower test and training errors and a high percentage of variance explained. However, this level of accuracy came at a cost since this model is not as interpretable compared to multiple linear regression since there are no parameters within to assist with interpretation. This model was still valuable since it indicated the most important variables through the variable importance plots.

The second best model was the logarithmic transformed multiple linear regression and there are some considerations that should be taken into account for it, especially since most of the model's assumptions were violated. The logarithmic transformation on the multiple linear regression model did not resolve the assumption violations, so other transformations such as Box-Cox transformation could have been applied. This transformation would have transformed the data to more closely represent a normal distribution. Furthermore, the model assumption violations in the errors having constant variance and the errors having uncorrelated error provided insight into what transformation was needed in order to properly fit the data, and that a more advanced regression model will have possibly been needed.

It is also important to note that there is a tradeoff in using the mean square error to decide which model is most appropriate for predicting housing prices. For instance, we decided not to choose the regression decision tree model in our final data analysis because the MSE was high, but this exclusion meant that we missed the opportunity to consider valuable information such as the most important variable that determines sales prices. In the mean square errors for each of the models that we applied in this project, the values were often high. Hence, it is important to note that standardizing the data could have improved the analysis since the models that we chose had many predictors and observations that increase the variability. Furthermore, since there were many variables in the logarithmic transformed multiple linear regression model, we could have considered reducing the number of predictors through further variable selection to simplify the model interpretation. In this same model, we primarily used the adjusted R^2, the p-values of the predictors, the training mean square error, and verification of model assumptions to assess the model's quality. However, the test error should have been considered to measure how well the model performs on houses that were not considered in the model's development.

As mentioned in the previous sections, we decided to exclude categorical variables to simplify model interpretation. However, this came at a cost since this meant we potentially lost valuable information on how the categorical variables influenced the sales prices. For further analysis, the categorical variables should have been taken into account by developing an interaction multiple interaction model or developing a kmeans clustering model to more accurately determine how these variables impacted sales price and the distribution of the data itself. We also applied the k-means clustering model, so that the multiple linear regression model can be applied to the resulting clusters. To improve the analysis, it would be helpful to further investigate the distribution of the clusters and use this information to determine how houses were grouped

References

House Prices - Advanced Regression Techniques (2022). *Kaggle*, https://www.kaggle.com/c/house-prices-advanced-regression-techniques

James, Witten, Hastie, Tibshirani. (2013). An Introduction to Statistical Learning with Applications in R. *Springer*, https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

Everitt, B. (2012). Cluster Analysis. Wiley.