Date: 12 – 13 - 2021

Housing Price Prediction on the Individual Sale of Home Prices in Ames, Iowa from January 2006 to July 2010

By: Kalide Endale, Donovan Johnson, Kevin Queally, Tanmay Shravge

Summary: This dataset is a collection of residential real estate sales from 2006 to 2010 collected by the Ames, Iowa Assessor Office. The dataset has a total of 2,930 observations and 80 variables. Out of the 80 variables, 20 are continuous, 14 are discrete, 23 are nominal, and 23 are ordinal. For the class final project, the objective was to determine which factors, qualitative or quantitative, are important features for the prediction of home sale prices. Furthermore, the team developed statistical models to predict home sale prices, specifically for homes in Ames, Iowa. Prior to analysis, exploratory data analysis was conducted to explore the data and discover patterns, spot anomalies, and present graphical representation that is beyond the formal modeling methods. Additionally, the dataset was cleaned and transformed for prediction. One of the challenges of working with this dataset, oddly, was also one of its strengths. The dataset has an extensive record of each sale. The large number of variables in the dataset (80) are quality and quantity attributes of each property. While this was a great problem to have, the main challenge was to reduce the number of variables necessary for predicting house sales without compromising prediction power or violating model assumptions. For analysis, the team utilized forward stepwise regression and random forest regression. Lastly, the team selected the final predictive model and compared the predicted home sale price with the actual sale price of each property. Ultimately, the best predictive model had six variables, which are: overall material finish and quality, above ground living area (measured by square feet), neighborhood, total basement square feet, year remodeled date, and year built. The most important predictor of sale price was above ground living area, and the second most important factor was the overall material finish and quality of the home. The root mean square error of the final model was 0.05681 (on a log10 scale) and mean absolute error was $17,948.90 (on actual dollar scale).

Methods: Initial first steps were to examine the data by teasing out some of the relationships that were present.

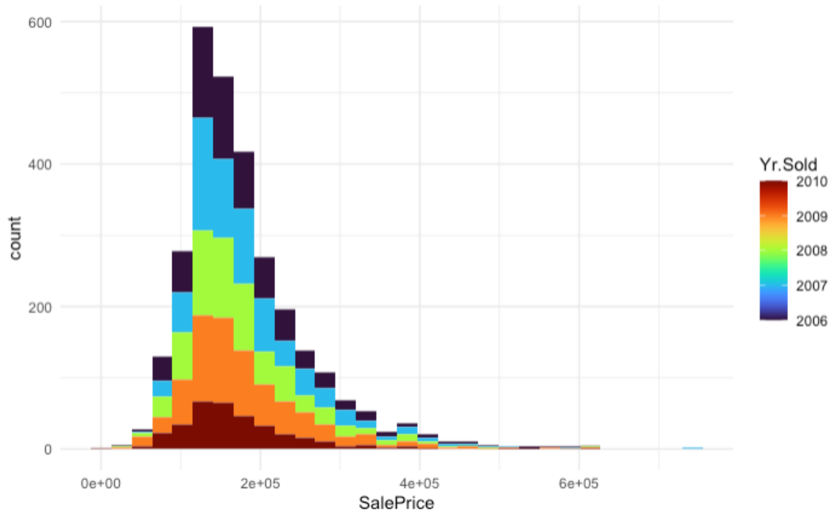
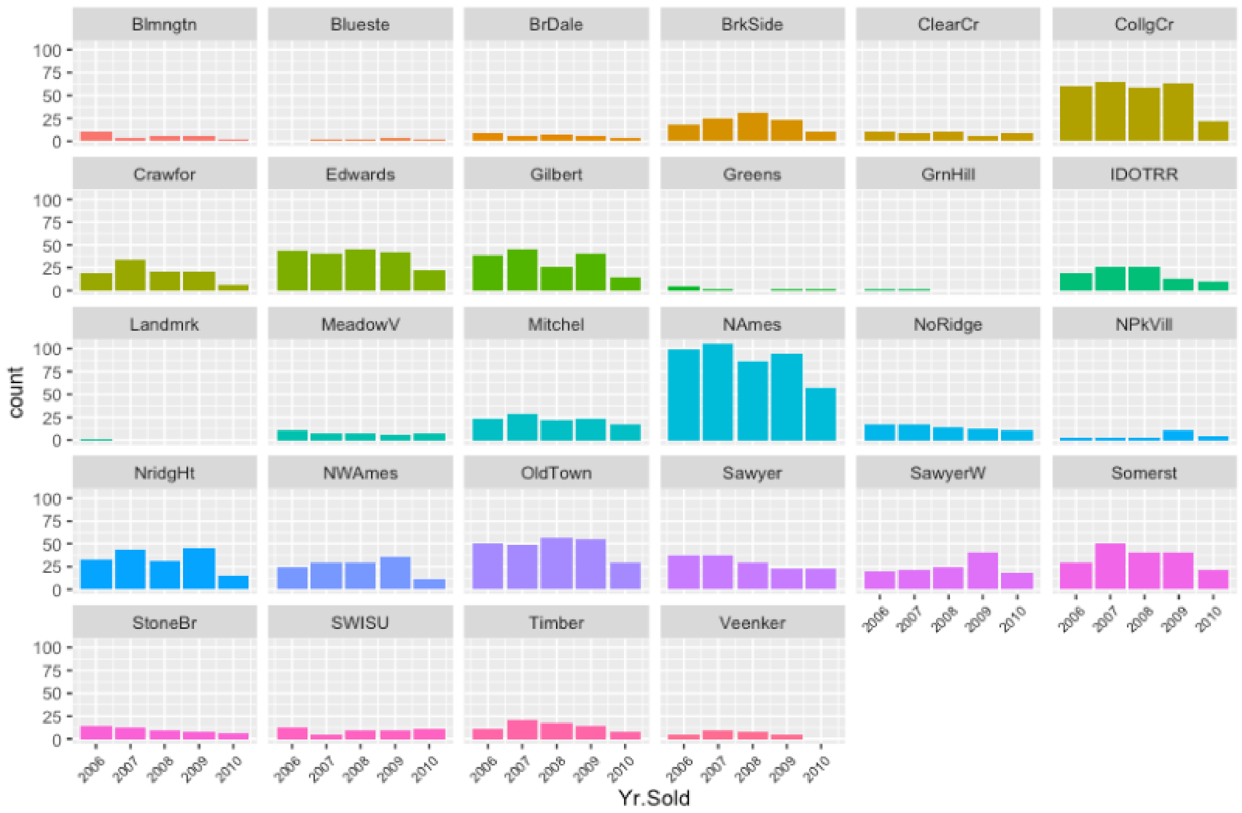


Figure 1

Figure 1 shows the distribution of sale price organized by year, with different colors representing the year the house was sold. From this histogram, it is evident that most of the homes in the dataset were sold in 2006, 2007, and 2008. The mean sale price is $180,796, with the maximum being $755,000 and the minimum being $12,789. To ensure these large outliers didn't negatively impact prediction ability, the team normalized the dataset by removing homes with a sales price greater than or less than three standard deviations from the mean. Furthermore, the general trend of the histogram is right skewed. Sale price has been 'logged' to symmetrize the distribution.



<u>Figure 2</u>

Figure 2 shows the distribution of homes sold per year, faceted by the neighborhood within Ames city. The neighborhoods "NAmes" and "CollegeCr" have the most houses sold in the period and "GrnHill", "Landmrk" have the least number of houses sold. Considering a portion of the project goal was to develop a predictive model for sale price, it's important to make sure there were sufficient data points for each neighborhood to partition the dataset into training (50%), validation (25%), and testing (25%) sets. Since "GrnHill" and "Landmrk" had fewer than 3 observations, these neighborhoods were removed from the list of factors.

After normalizing the distribution of sale price and removing neighborhoods with limited observations, the next step was reducing the number of candidate predictor variables. The first way this was accomplished was by using contextual knowledge to eliminate variables that were either overlapping with other variables in the dataset or unnecessary for prediction. For example, the dataset had variables for: the total square feet above the ground floor, the square feet of the first floor, and the square feet of the second floor. To reduce multicollinearity, which would negatively affect the final model, only total square feet above the ground floor was kept as it encompasses both other variables. Another example is the area of a garage and the number of cars a garage can hold as variables. For prediction, it is sufficient to keep the number of cars a garage can hold and drop garage area since these two variables encode similar information. This pre-processing step reduced the total observations to 2,657 and the number of predictor variables to 43.

To further reduce the number of variables, ANOVA was then used as an independent test for statistical significance. Significance for numeric variables was tested based on the fit of a linear line-of-best-fit, whereas categorical variables were tested using one-way ANOVA. This analysis was conducted to ensure that the variables included in the final model were significant predictors of sale price. Since nearly 50 variables were being tested, a Bonferroni correction on p-values was utilized. This correction limits the number of Type I errors, while assuming the risk that some variables that are significant will be removed. This risk felt appropriate due to the large number of variables at hand. After running this analysis with an alpha cutoff of .05, only 22 variables were allowed to pass for further analysis.

Features <chr>	**Correlation** <dbl>	**Absolute Value of Correlation** <dbl>
SalePrice	1.000000000	1.000000000
Overall.Qual	0.810765090	0.810765090
Gr.Liv.Area	0.739480699	0.739480699
Garage.Cars	0.661374994	0.661374994
Total.Bsmt.SF	0.654681272	0.654681272
Year.Built	0.567250250	0.567250250
Year.Remod.Add	0.552944564	0.552944564
Neighborhood	0.335041380	0.335041380
Lot.Area	0.328760305	0.328760305
Bsmt.Unf.SF	0.163236558	0.163236558
Bsmt.Qual	0.119118528	0.119118528
Kitchen.Qual	0.110993057	0.110993057
Foundation	0.100697938	0.100697938
Exter.Qual	0.096126794	0.096126794
MS.SubClass	-0.075598534	0.075598534
Sale.Condition	0.074733650	0.074733650
Sale.Type	0.072511842	0.072511842

Figure 3

Now that only significant variables were being considered, the next investigative step was to find out which predictor variables have the highest correlation with sale price. For numeric variables, Pearson correlation coefficient was used to quantify the relationship, whereas categorical variables were tested with Theil's U correlation coefficient. Theil's U is a measure of the predictive ability of discrete variables and its outputs were compared directly with Pearson correlations. Figure 3 displays some of the outputs of these methods, ordered by the absolute value of their correlations. After the first eight variables (in green), there was a significant correlation drop. Variables with correlations below that of Lot.Area were removed, leaving eight variables to be included for the model selection.

Model Selection: Random Forest is an algorithm which can be used both for classification and regression. Random forest models are constructed by using a collection of decision trees based on the training data. Instead of taking the target value from a single tree, the random forest algorithm makes a prediction on the average prediction of a collection of trees, which helps reduce the overfitting that can occur when using single decision trees.

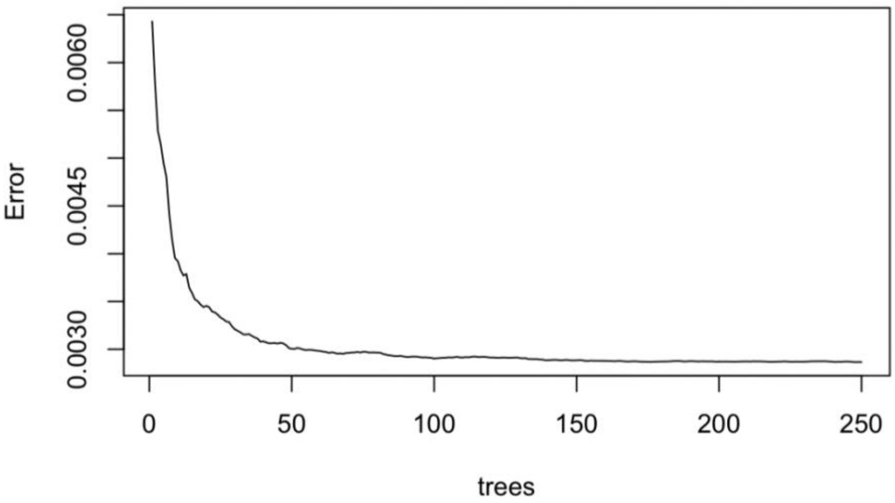


Figure 4

Random forest regression was performed using the final eight predictor variables from Figure 3. Figure 4 displays the relationship between the mean of squared residuals (Error) and the number of trees in the random forest, which reached 0.002865. The root mean squared error (RMSE) is a predictive accuracy metric which is calculated using the difference between actual vs predicted sales price. RMSE was used instead of other model diagnostics because it is sensitive to large predictive discrepancies. For model selection RMSE was calculated on the validation set, resulting in 0.05668.

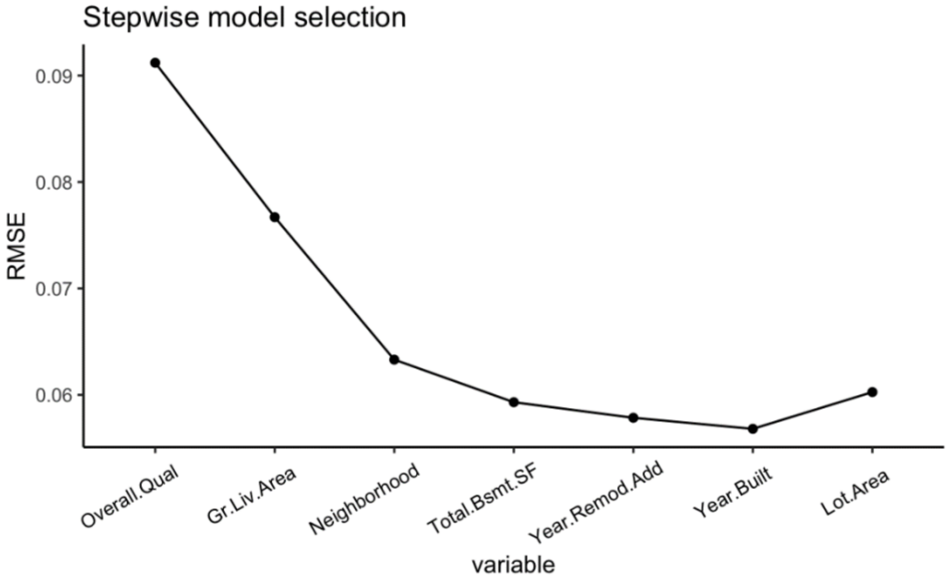


Figure 5

After creating this model, another approach was attempted using stepwise linear regression. Forward stepwise selection is a variable selection method which begins with a model that contains no variables (called the null model) then starts adding the most significant variables one after the other. In this model, new variables were added to the model until the RMSE on the validation set increased or stopped decreasing. Using this method, variables were added in the order of: overall quality, above ground living area, neighborhood, total basement area, year the house was most recently remodeled, and finally the year the house was built. Figure 5 demonstrates that, after adding the next best predictor variable, Lot Area, the RMSE increased. This is the stopping point for the stepwise selection, and only 6 variables were included in this final model while the other 2 were dropped. Final RMSE on the validation set was 0.05681, which is a 0.00013 increase from the random forest model.

Even though the random forest model slightly outperformed the stepwise linear regression in terms of predictive ability, the stepwise linear regression was selected as the final model. This is because random forest is significantly more computationally intensive than linear regression, which results in longer run times. Random forest is also a 'black box' in terms of prediction, as in there is little interpretation of variable importance or reasoning. Because our goal is to predict housing prices and determine the most important factors for home valuation, this small difference in prediction accuracy was outweighed by model interpretability and computational efficiency. Residual plots for this model are shown in the appendix.

Results: The final model selected was a stepwise linear regression that had six predictor variables as the main determinants of house sale price in Ames, Iowa. This model was selected due to its computational efficiency, ease of interpretability, and predictive accuracy.

Predicted Sale Price <chr>	**Actual Sale Price** <dbl>	**Error** <dbl>
$181,842.84	$195,500	$13,657.16
$234,586.03	$213,500	$21,086.03
$175,744.94	$175,900	$155.06
$170,528.05	$185,000	$14,471.95
$234,373.43	$212,000	$22,373.43
$96,066.28	$105,500	$9,433.71
$239,589.14	$220,000	$19,589.13
$322,650.80	$320,000	$2,650.80
$315,740.51	$319,900	$4,159.48
$175,697.59	$160,000	$15,697.58

Figure 6

The RMSE for the stepwise linear regression model on the test set was .05443, which was lower than the tests on the validation set. Figure 6 demonstrates the predictive accuracy of the stepwise model on a sample from the test set. Predicted sale price was compared to the actual sale price, with Error representing the absolute difference between the two. The mean absolute error (MAE), a measure of the average absolute price discrepancy between model predictions and actual sales, was $17,948.90 for the test set. Considering house prices are highly negotiable and the mean actual sale price was $180,796, this estimate error is within an acceptable range. The mean bias error (MBE), an estimate of the average model bias in the predicted sale price vs actual sale price, was -1,812.10, meaning the final model underestimated the final sales price by $1,812.10 on average. This indicates that further investigation may be needed to understand why predicted values were, on average, lower than actual values; however, this bias is still relatively low considering that sale prices are in the hundreds of thousands of dollars.

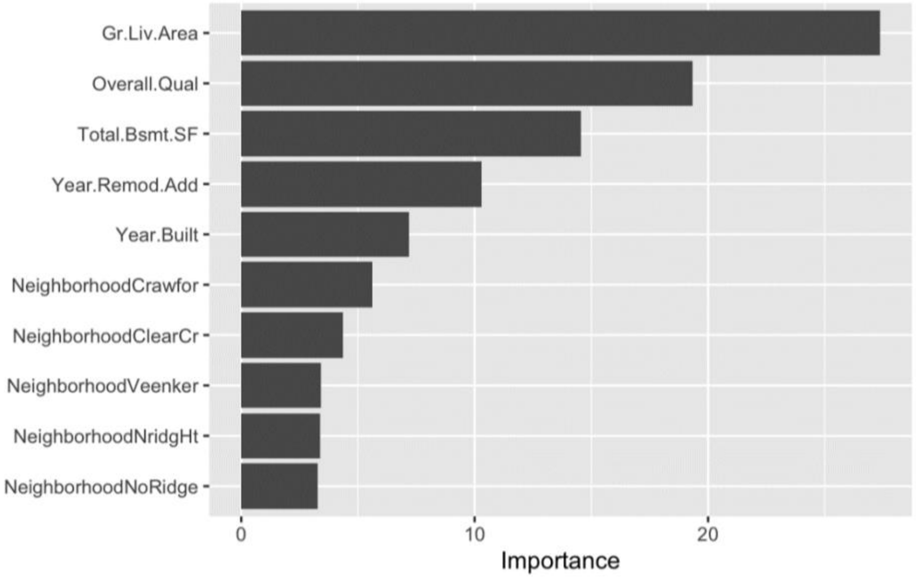


Figure 7

Examining model diagnostics more closely revealed that every predictor numeric variable was positively correlated with house sale price. Most neighborhoods either had a positive or insignificant effect, except for three neighborhoods that had negative correlations: Briardale Square, Northpark Villa, and Meadow Village. Upon further research, it's difficult to say why Meadow Village had a negative effect on sale price, however, the team did find information about the composition of homes in the other two neighborhoods. In Briardale Square, 60% of homes are rentals in this area, and on average have 37% lower expenses than the national average. Additionally, Northpark Villa is located near a college which could make the neighborhood an undesirable location for families to relocate to. Although there is not much information about Meadow Village, the reason for the negative correlation with house sale price is most likely a similar socioeconomic factor that is having an exogenous effect on the model. Figure 7 demonstrates the relative significance or importance of each variable in the final model. This ordering is slightly different from the ordering implied by the stepwise selection in Figure 5. This discrepancy is likely caused by variables that are interrelated, which the stepwise selection (Figure 5) accounts for, while importance testing (Figure 7) does not.

With the elimination of 74 variables from the dataset, the final linear model is able to predict the sale price of houses in Ames, Iowa from 2006 to 2010 with just 2 continuous variables, 3 discrete variables, and 1 nominal variable.

Discussion: In this report, the team analyzed the relationship between house sale price and various explanatory variables. The six most important factors for home price valuation in Ames, Iowa were: the overall material and finish quality, the above ground living area in square feet, the total square feet of the basement, the neighborhood within Ames, the year the house was most recently remodeled, and finally, the year the house was built. This means that, when evaluating a home – either for purchase or for sale – that these factors will be most correlated with a higher sales price. This model is important for home buyers and sellers, real estate agents, or real estate software companies (e.g., Zillow) for the purpose of home valuation in the Ames, Iowa area. With this model, stakeholders can make better decisions when putting a house on the market or making an offer, as they will have an objective valuation metric of a home.

While this model can generally predict the price range of a home, there are many limitations and potential improvements. One such limitation is that the data for this model is from 2006-2010, meaning that the dataset's modern-day predictive ability is unknown. Given that house prices have dramatically increased in recent years, it is likely that this model or dataset is no longer relevant to the market and would need to be trained on new data to predict house prices today. Another limitation of this model is that it fails to consider other important factors, many of which are immeasurable qualities, to evaluate a home. This, combined with the fact that outliers outside of three standard deviations from the mean sales price were removed, could potentially cause the current model to fail when predicting the value of rare or unique houses. Most importantly, an underestimation bias was detected in the final model. Although this bias was not extreme, it is still important to investigate its cause so that it may be prevented in the future.

This model could be enhanced in several different ways. For one, ANOVA was used as a dimension reduction technique, which is not the best method when the end goal is predictive power. While this method did aid in reducing the number of candidate predictor variables, further testing on correlation would have been better suited for this task. Correspondingly, the Bonferroni correction during the ANOVA process may have caused exorbitant eliminations since it is a conservative correction against Type I errors. Thus, through our ANOVA testing, many predictor variables which may have led to better predictions, could have been removed on account statistical insignificance. This was not the best approach, and future work should focus on correlations with the sale price alone. Another potential improvement would be through more contextual knowledge of home valuation. Many variables were eliminated through contextual understanding of the problem; however, no consultation was done with subject matter experts, such as real estate agents. It is possible that some variables that were removed were, in fact, better suited for the model. Finally, no unsupervised dimension reduction techniques were attempted, due to the mix of categorical and numerical variables in the dataset. While dimension reduction may have aided in reducing multicollinearity, the methods required to conduct this type of dimension reduction were outside the scope of this project.

Statement of Contributions:

Kalide Endale: Contributed with proposal write-up, data cleaning, variable reduction, data analysis, report write up, exploratory data analysis

Donovan Johnson: Contributed with proposal write-up, data cleaning, variable reduction, data analysis, report write up, ran ANOVA on all potential predictors.

Kevin Queally: Contributed with data partitioning, ANOVA and subsequent corrections, correlation analysis (Pearson & Theil's U), random forest regression, stepwise regression, model diagnostics, report write up.

Tanmay Shravge: Contributed with data cleaning, correlation analysis for the categorical variables (Theil's U), variable reduction, tuning random forest regression, analyzing the predictors for random forest and stepwise regression, in the presentation part, and report write up.

Date: 12 – 13 - 2021

Work Cited:

De Cock, Dean. “Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project.” Journal of Statistics Education, vol. 19, no. 3, 2011, https://doi.org/10.1080/10691898.2011.11889627.

De Cock, Dean. “Overview.” Data Sets, https://www.openintro.org/book/statdata/?data=ames.

De Cock, Dean. “Ames Iowa: Alternative to the Boston Housing Data Set.” *Jse.amstat.org*, http://jse.amstat.org/v19n3/decock/DataDocumentation.txt.

Liu, Ziqiao. “Insights on Housing Data: Multiple Factors behind House Price.” Data Science Blog, 25 Oct. 2016, https://nycdatascience.com/blog/student-works/key-insights-ames-iowa-housing-data-multiple-factors-behind-house-price/.

Appendix:

Section 1: Continued EDA

With the processed and transformed dataset, the correlated variables were plotted against 'logged' sale price. When overall material and finish quality of home is plotted against the log10 of sale price (Figure A), there is a positive linear relationship.

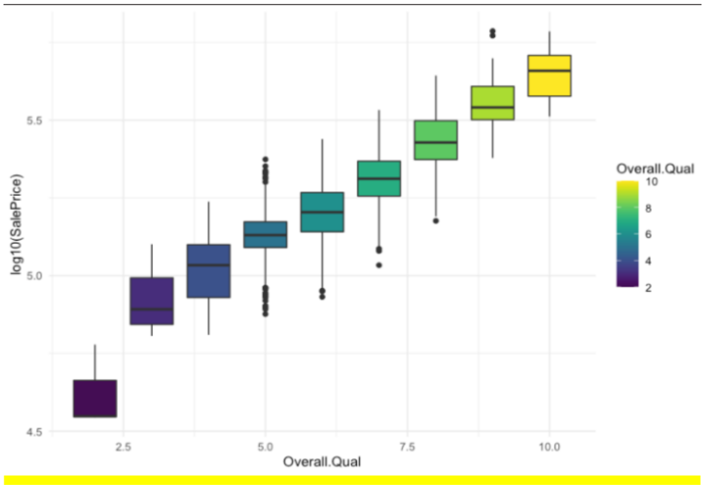


Figure A

Additionally, the same positive linear relationship is present when plotting above ground living area to average log10 sale price (Figure B).

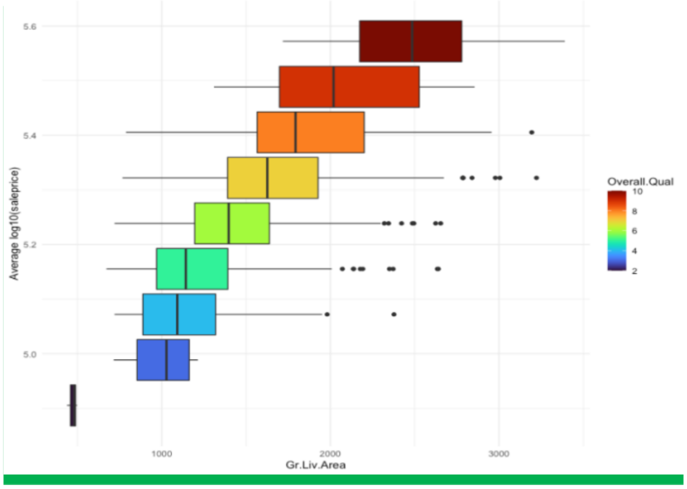


Figure B

With garage cars as a predictor variable of sale price, the price of a home increases with an additional increase in the number of garage space available for cars, but the price increase peaks at three and tapers down after (Figure C). This suggests that if a home has more than three car spaces, each additional increase in garage space would not increase the sale price of the home.

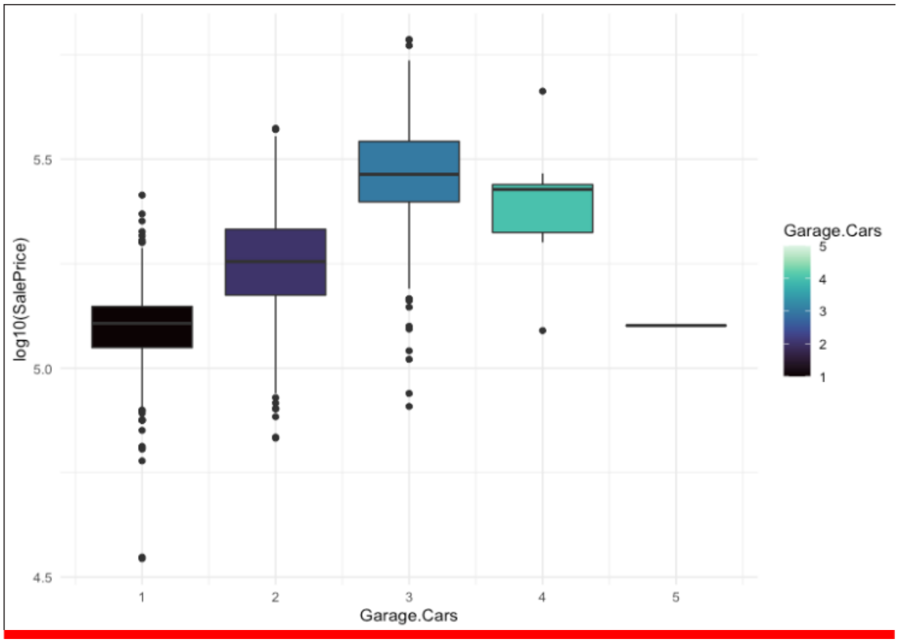
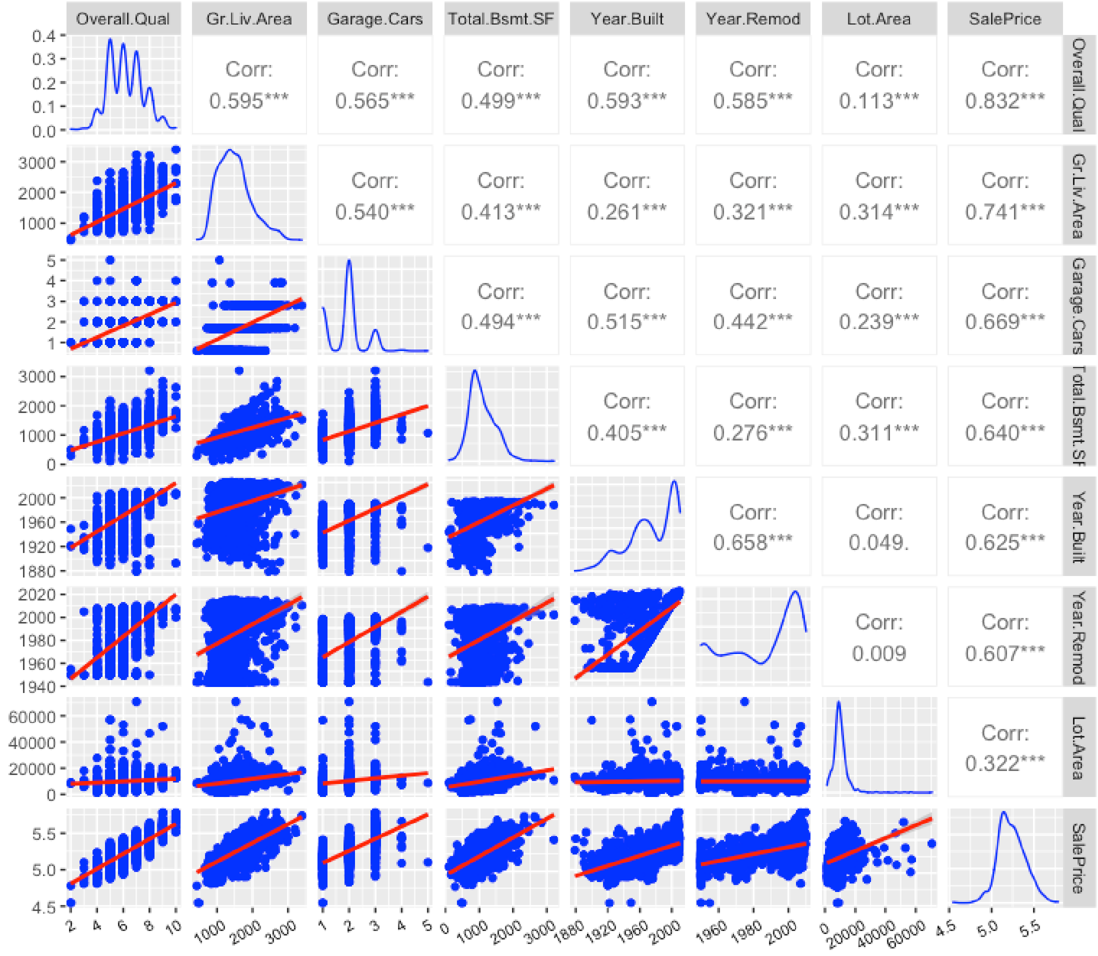


Figure C

Section 2: Variable Selection & Multicollinearity

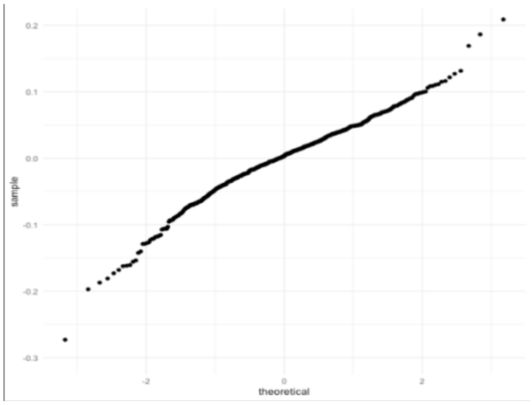
Correlation and Distribution plots of the final numeric variables

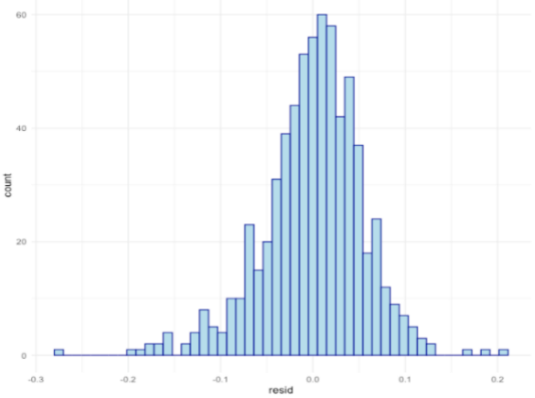


Date: 12 – 13 - 2021

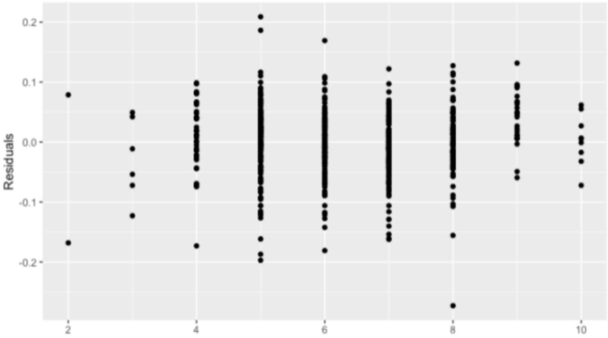
Section 3: Model Assumptions & Diagnostics

Model Assumption intact, qq-plot looks linear, the residual histogram is randomly distributed, and the when the residual of the model are plotted against the variables in the final model the errors seem to be random.

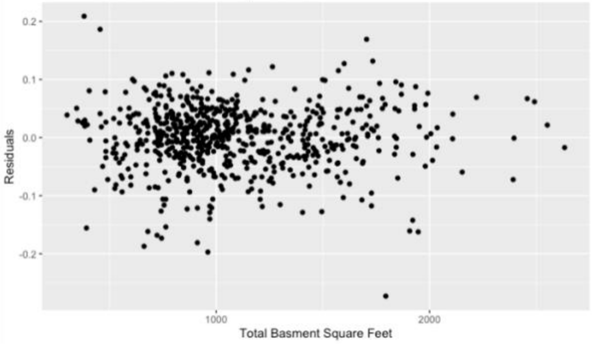




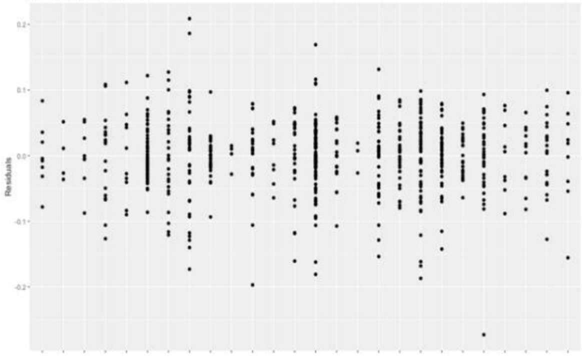
Residuals vs. Overall Quality



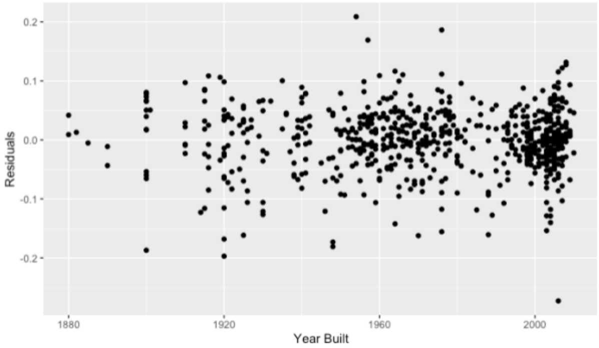
Residuals vs. Total Basment Square Feet



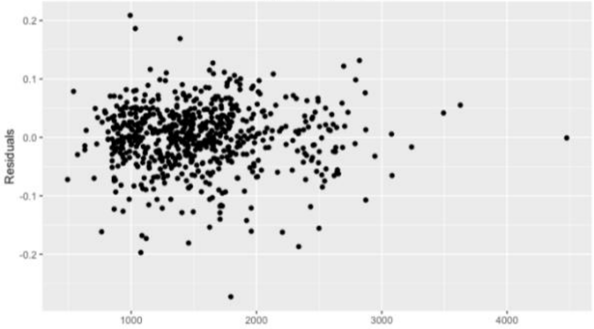
Residuals vs. Neighborhood



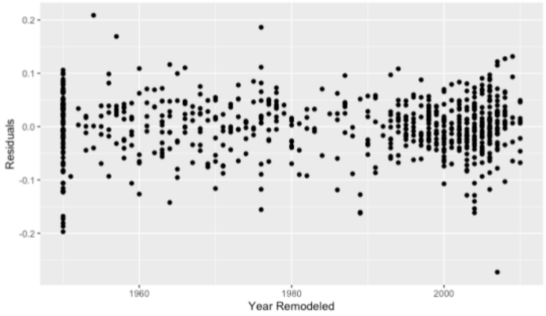
Residuals vs. Year Built



Residuals vs. Above Ground Living Area Square Feet



Residuals vs. Year Remodeled


Residuals
Year Remodeled