Decision 518

Team 01: Anika Abrahamson, Yaqiong (Juno) Cao, Dexter Nguyen, Michael Ruch, Xinying (Silvia) Sun

Creating a Linear Regression Model to Help Predict Optimal Sales Price for Homes in Ames, Iowa

September 26, 2020

Contents

Business Understanding

Within the housing market, a *Comparative Market Analysis*, or CMA, is utilized by the broker to present the seller with a proposed sale price and a comprehensive justification for this price (Miller 2018). Although many brokers utilize software to complete a CMA, personal experience and intuition are also employed to decide the proposed price. The goal of our analysis is to create a regression model for pricing homes in the Ames, Iowa housing market. In theory, pricing homes closer to their "real value" (based on a concrete model) will result in lower resource use on the part of the broker/agency, and thus, a quicker (and more lucrative) sale. There are various models that websites, for example, Zillow.com, apply to provide estimates on the market value of a particular home (McDonald 2006). Our analysis will build a model specifically for Ames, Iowa that real estate brokers can utilize for their CMA reports to be more confident in their proposed price.

Data Understanding

Our analysis will use the dataset compiled by Dean DeCock in 2011 and published on Kaggle by Mehdi in 2018. The dataset contains 80 variables, which were recorded for 2930 properties in Ames Iowa (DeCock 2011). This data will allow us to create a linear regression model to find how different independent variables affect our dependent variable, sales price. The knowledge of how each variable will impact the home's price will help real estate brokers better assess a proper sales price for a home in Ames, Iowa.

Data Preparation

Data Cleaning

Our first step was to clean and prepare the data for analysis. We removed the extraneous columns of "Order" and "PID" because they were irrelevant to our research. We chose to change the subclass from numerical to categorical to simplify the computation and visualization of correlation. We removed all N/A values and used a sampling method to impute the missing values, where appropriate (Buuren & Groothuis-Oudshoorn 2011). We also removed columns with over 85% of values missing. In some cases, we replaced all missing categorical values with the modal value. For our categorical variables, we created dummy variables to allow for numerical calculation of correlations. With two categorical variables, Kitchen_Qual and Bsmt_Qual, we replaced the abbreviated name with a more intuitive one and reordered the factor levels as Ex = Excellent, Gd = Good, Fa = Fair, Ta = Typical, Po = Poor.

Data Exploration and Transformation

To see which variables are likely to affect the price of homes in Ames, IA the most, we ran a correlation analysis of our independent variables against our dependent variable, sale price. Once this was completed, we chose to keep the top 10 variables of interest, which had the highest correlation with price (Figure 1). In order of highest correlation, these variables are:

1. Overall.Qual: Overall material and finish quality
2. Gr.Liv.Area: Above grade (ground) living area square feet
3. Garage.Cars: Size of garage in car capacity

4. Garage.Area: Size of garage in square feet
5. X1st.Flr.SF: First Floor in square feet
6. Total.Bsmt.SF: Total square feet of basement area
7. Bsmt.Qual_Excellent: Basement quality at Excellent
8. Full.Bath: Full bathrooms above grade
9. Kitchen.Qual_Excellent: Kitchen quality at Excellent
10. Year.Built: Original construction date

Looking at the distribution of our dependent variable SalePrice, we concluded that our data is not normally distributed (Figure 2). After running a QQ plot, it was clear that we needed to transform our data so it would be normally distributed (Figure 3). The new QQ plot demonstrates that our log(SalePrice) values are much more normally distributed, allowing us to move forward with our analysis (Figure 4).

Next, we plotted the marginal distributions of the key categorical variables of interest and displayed their relationship with price (Figures 5 and 6). Not surprisingly, we found clear positive correlations between kitchen quality and price and between basement quality and price. For numerical variables, the first step to further analyze the relationship with our dependent variable was to create density plots visualizing the spread of the data. After analyzing the density plots, we plotted the interaction between our numeric variables of interest and our dependent variable of price. The variables ground floor, living area, garage area, first-floor square footage, total basement square footage, and year built show a similar pattern to the overall quality graph (Figure 7).

We found a few correlations that did not play out as expected. When inspecting the variables garage cars and full bath (and their respective relationships with price - Figures 8 and

9), we saw the expected positive correlation up to a point, but then a decline. Homes with four-car garages, for example, had generally lower prices than those with three-car garages. Similarly, homes with four full baths had generally lower prices than those with three full baths. We assume that fewer buyers are looking for these attributes and, as such, four-car garages simply command a lower price. It is also likely that there are fewer of these homes on the market and that other variables simply have a more significant impact on price across the relatively small sample of four-car garage / four full bath homes.

After examining our independent variables of interest, we decided boxcox transformations were needed in order to make sure the data was normal and made sense when using it alongside our log(SalePrice). After calculating the lambda values and plotting our transformed, normalized data with our transformed, normalized price, only five of our initial ten variables of interest needed to be normalized with the logarithmic transformation. These five are Gr.Liv.Area, X1st.Flr.SF, Garage.Cars, Garage.Area, and Total.Bsmt.SF. The comparisons between the transformed and not transformed variables are shown in figures 10-14.

Last, we considered if the collinearity problem existed in our analysis. Calculating different correlations among our independent variables, we discovered the high correlation (0.8461) between two variables: Garage.Cars and Garage.Area. This, in addition to the output from our previous analysis, was the impetus to remove Garage.Cars from our regression models.

Modeling

From our exploratory data analysis, we know that sale price is highly correlated with a number of variables (our "top 9"). We will now employ the use of linear regression to build an optimal pricing model for homes in this market.

Selected independent variables

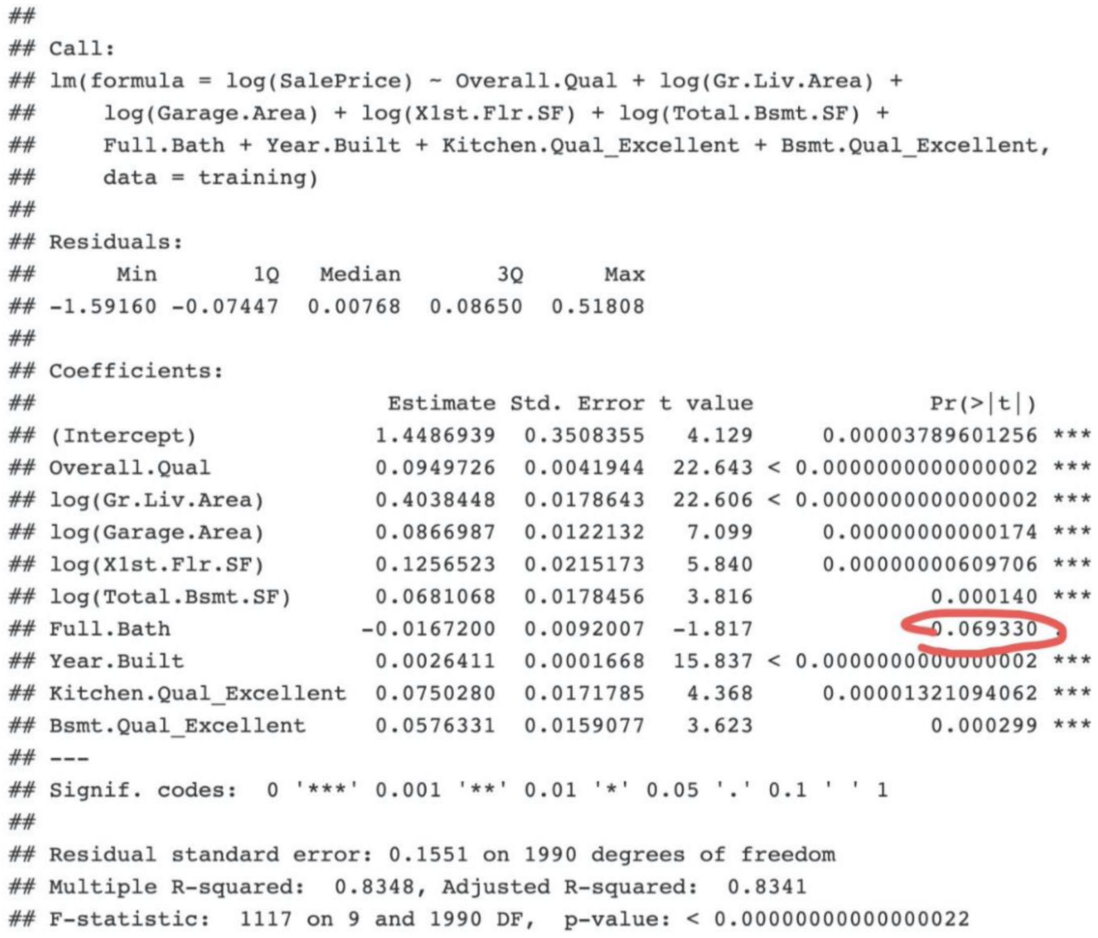
As a result of our exploratory data analysis, we selected nine independent variables for use in our regression modelling, 4 of which were logarithmically transformed: Overall.Qual, log(Gr.Liv.Area), log(Garage.Area), log(X1st.Flr.SF), log(Total.Bsmt.SF), Bsmt.Qual_Excellent, Full.Bath, Kitchen.Qual_Excellent, and Year.Built.

Regression Modelling:

We created four multi-regression models. Our first model (“Model 1”) included our “top 9” explanatory variables (variables highly correlated with price). We developed our second model (“Model 2”) by removing the Full.Bath variable, which was not statistically significant with its p-value > 0.05. For comparison, we developed Model 3, which included all explanatory variables in the original dataset. In Model 4, we removed non-significant variables (p-values > 0.05) from this more comprehensive model.

Based on the p-value and R-squared performance (and not wanting to over-fit the model), we picked Model 2, which includes all the significant independent variables without the high associated p-values, as our final model. The summaries of these models are below.

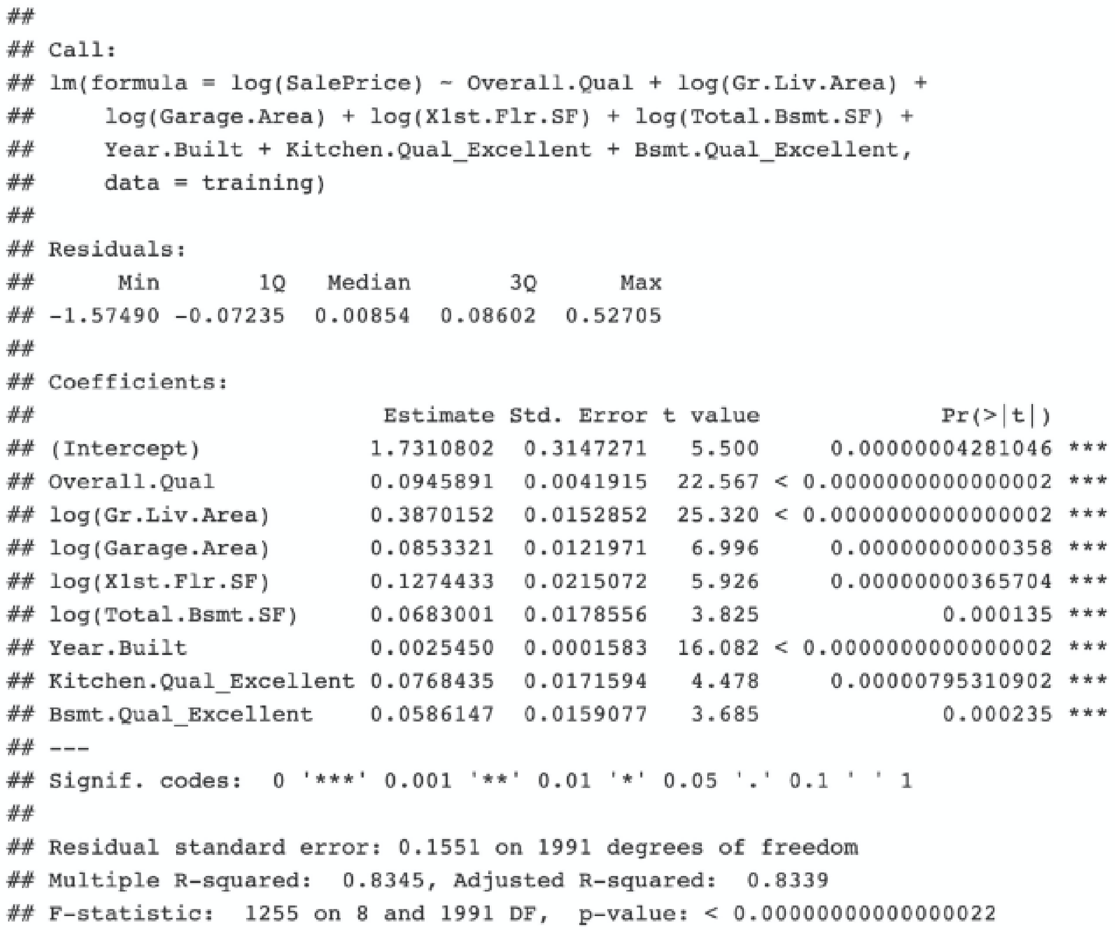
Model 1 (Top 9 explanatory variables) summary:



```
##
## Call:
## lm(formula = log(SalePrice) ~ Overall.Qual + log(Gr.Liv.Area) +
##     log(Garage.Area) + log(X1st.Flr.SF) + log(Total.Bsmt.SF) +
##     Full.Bath + Year.Built + Kitchen.Qual_Excellent + Bsmt.Qual_Excellent,
##     data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.59160 -0.07447  0.00768  0.08650  0.51808
##
## Coefficients:
##                          Estimate Std. Error t value             Pr(>|t|)
## (Intercept)             1.4486939  0.3508355   4.129     0.00003789601256 ***
## Overall.Qual            0.0949726  0.0041944  22.643 < 0.0000000000000002 ***
## log(Gr.Liv.Area)        0.4038448  0.0178643  22.606 < 0.0000000000000002 ***
## log(Garage.Area)        0.0866987  0.0122132   7.099     0.00000000000174 ***
## log(X1st.Flr.SF)        0.1256523  0.0215173   5.840     0.00000000609706 ***
## log(Total.Bsmt.SF)      0.0681068  0.0178456   3.816             0.000140 ***
## Full.Bath              -0.0167200  0.0092007  -1.817             0.069330 .
## Year.Built              0.0026411  0.0001668  15.837 < 0.0000000000000002 ***
## Kitchen.Qual_Excellent  0.0750280  0.0171785   4.368     0.00001321094062 ***
## Bsmt.Qual_Excellent     0.0576331  0.0159077   3.623             0.000299 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1551 on 1990 degrees of freedom
## Multiple R-squared:  0.8348, Adjusted R-squared:  0.8341
## F-statistic:  1117 on 9 and 1990 DF,  p-value: < 0.00000000000000022
```

The p-value of Full.Bath is 0.069330. This value is too high, so we removed this variable, yielding our second model.

Model 2 (Top 8 explanatory variables) summary:

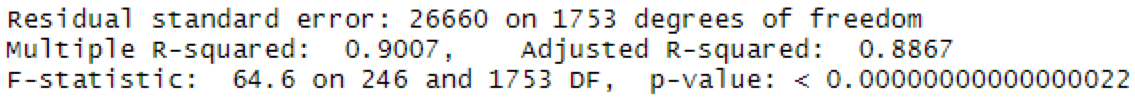


```
##
## Call:
## lm(formula = log(SalePrice) ~ Overall.Qual + log(Gr.Liv.Area) +
##     log(Garage.Area) + log(X1st.Flr.SF) + log(Total.Bsmt.SF) +
##     Year.Built + Kitchen.Qual_Excellent + Bsmt.Qual_Excellent,
##     data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57490 -0.07235  0.00854  0.08602  0.52705
##
## Coefficients:
##                         Estimate Std. Error t value             Pr(>|t|)
## (Intercept)            1.7310802  0.3147271   5.500     0.00000004281046 ***
## Overall.Qual           0.0945891  0.0041915  22.567 < 0.0000000000000002 ***
## log(Gr.Liv.Area)       0.3870152  0.0152852  25.320 < 0.0000000000000002 ***
## log(Garage.Area)       0.0853321  0.0121971   6.996     0.00000000000358 ***
## log(X1st.Flr.SF)       0.1274433  0.0215072   5.926     0.00000000365704 ***
## log(Total.Bsmt.SF)     0.0683001  0.0178556   3.825             0.000135 ***
## Year.Built             0.0025450  0.0001583  16.082 < 0.0000000000000002 ***
## Kitchen.Qual_Excellent 0.0768435  0.0171594   4.478     0.00000795310902 ***
## Bsmt.Qual_Excellent    0.0586147  0.0159077   3.685             0.000235 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1551 on 1991 degrees of freedom
## Multiple R-squared:  0.8345, Adjusted R-squared:  0.8339
## F-statistic:  1255 on 8 and 1991 DF,  p-value: < 0.00000000000000022
```

In Model 2, all identified variables are highly correlated with our target variable (SalePrice) and show statistical significance. All the variables have a positive relationship with SalePrice. To be more specific, we expect an average increase of 0.4% in SalePrice for every 1% increase in Gr. Liv. Area, holding other variables constant. Another interpretation can be seen from two variables: Kitchen.Qual and Bsmt.Qual. These two only have positive impacts on the sale price if the value is ‘Excellent’. To dive deeper into the regression analysis, we also tried to find interaction among independent variables, but found no meaningful insights.

Aside from the first two models, for which we used our “top 9” and “top 8” variables (determined through EDA), we would like to have a more comprehensive model with all variables included to see how well it performs. Model 3 is this comprehensive model. It is obvious that the R-squared is higher than the second model, since the more variables we input, the more accurate our model will be. However, such a model is clearly over-fit.

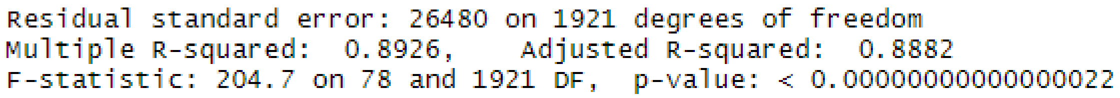
Model 3 (All independent variables) summary:



```
Residual standard error: 26660 on 1753 degrees of freedom
Multiple R-squared:  0.9007,    Adjusted R-squared:  0.8867
F-statistic:  64.6 on 246 and 1753 DF,  p-value: < 0.00000000000000022
```

Examining the p-values of each variable in Model 3, we found that many variables had large p-values. We dropped those high p-value variables, yielding our fourth and final model. The performance of Model 4 improved a little bit compared to Model 3, with an R-squared value of 0.89 (0.05 higher than Model 2). However, this model still includes over 100 variables.

Model 4 summary:



```
Residual standard error: 26480 on 1921 degrees of freedom
Multiple R-squared:  0.8926,    Adjusted R-squared:  0.8882
F-statistic: 204.7 on 78 and 1921 DF,  p-value: < 0.00000000000000022
```

We ultimately chose Model 2, which includes the “top 8” highly correlated variables, as our final model. Our P-values indicate that each variable in our top-8 model is of statistical significance. The R-squared is relatively strong, with a value of 0.84. Nevertheless, the R-squared value of Model 4, our more comprehensive model, is higher still. Despite Model 4’s marginally higher R-squared value, from a practical perspective, Model 2, with only eight independent variables, is more efficient in terms of resources used to collect and process data. Model statistical performance is not the only standard to determine the best model, however.

Model Evaluation

After running our two models: Model 1 and Model 2, we used R-squared and AIC to evaluate our model performance. As we expected, Model 2 is best suited for our business use case. We compared the R-Squared and AIC for Model 1 and Model 2. It is obvious that the evaluation factor is quite close in the two models. The model performance remains high after removing the high p-value variable in the first model. Since we used fewer variables to predict SalePrice and didn't hurt the model performance after removing said high-p-value-variable, we determine that Model 2, which includes highly correlated variables without large p-values, is the highest performance model we have so far.

	AIC	R-Squared
Model 1 - High Correlated Variables with large P-value	-1856.32	0.8388
Model 2 - High Correlated Variables w/o large P-value	-1856.56	0.8386

Implications, limitations, and conclusion

By analyzing the data collected in the Ames, Iowa real estate market, we were able to create a model that can help future sellers price their homes in the market to sell quickly while still generating a profit. The most important factors when determining price, as determined by our analysis, are the year built, excellent kitchen and basement quality, the square footage of both the basement and first floor, the square footage of both above-grade living area and garage area, and the overall quality (as determined by material and finish) of the home. Because our model is based on these variables, we believe it to be a useful tool for real estate agents to utilize in the Ames, Iowa market.

However, since we used the 2011 dataset to build the model and the real estate market is constantly changing, our best model might not fit the current market. Going forward, we would recommend frequently recording housing specs and sales prices in the Ames area and maintaining a database with the relevant information to continually improve on the model's ability to predict sales price, even in the face of an ever-changing market landscape. Aside from considering the time-efficiency of using our model, we have to consider a better way to deal with missing values. In addition, regression modelling has its limitations. As such, we would recommend exploring more comprehensive machine learning models and different evaluation methods in the future.

Appendix: Referenced Figures

Figure 1. Correlation Matrix

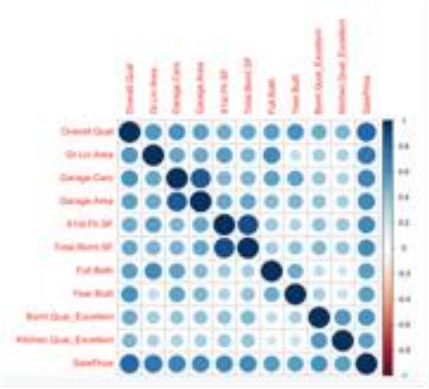


Figure 2.

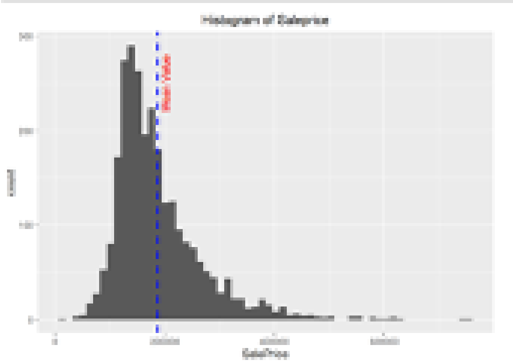


Figure 3.

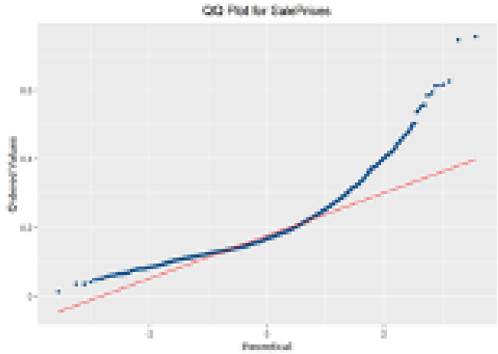


Figure 4.

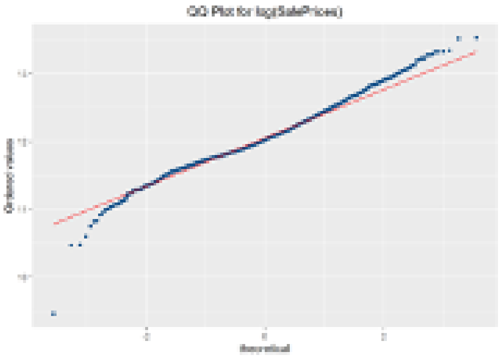


Figure 5. Sale Price vs. Basement Quality

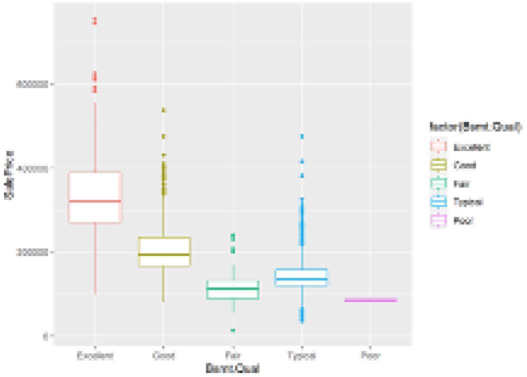


Figure 6. Sale Price vs. Kitchen Quality

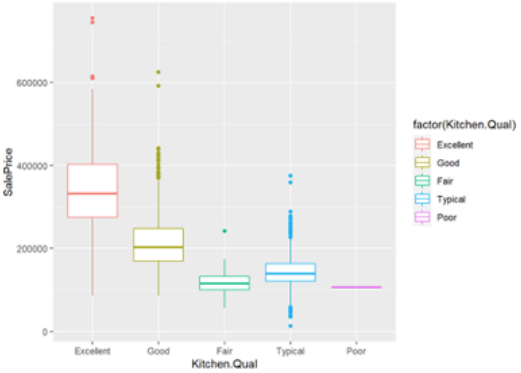


Figure 7. Overall Quality vs. Sales Price

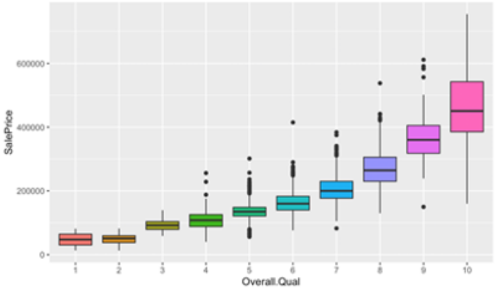


Figure 8. Garage Cars vs. Sales Price

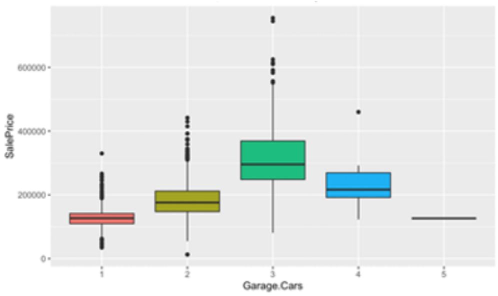


Figure 9. Full Baths vs. Sale Price

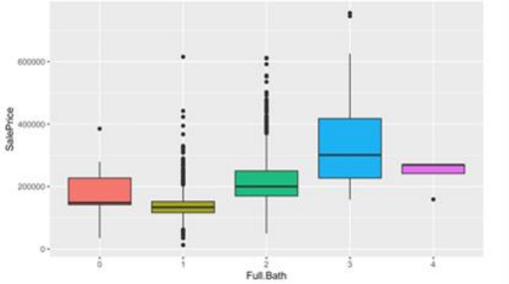
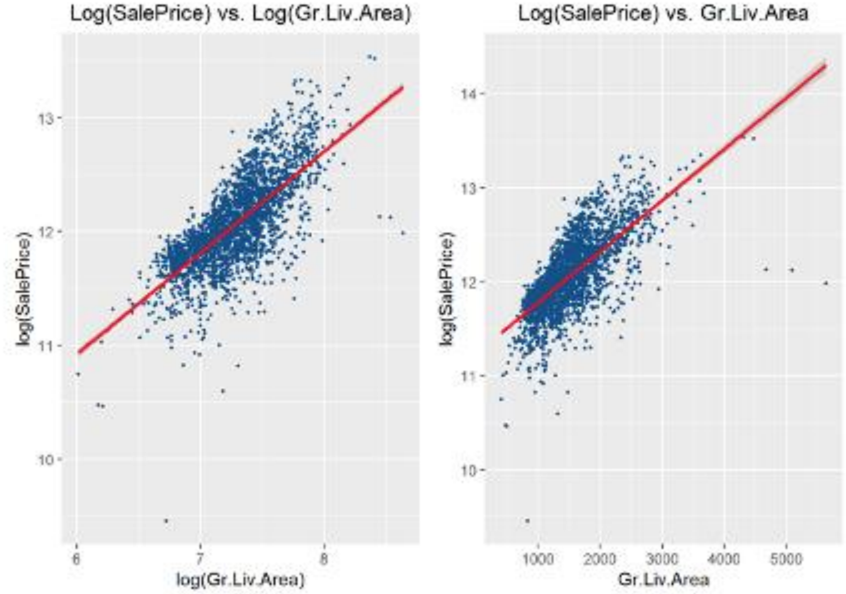
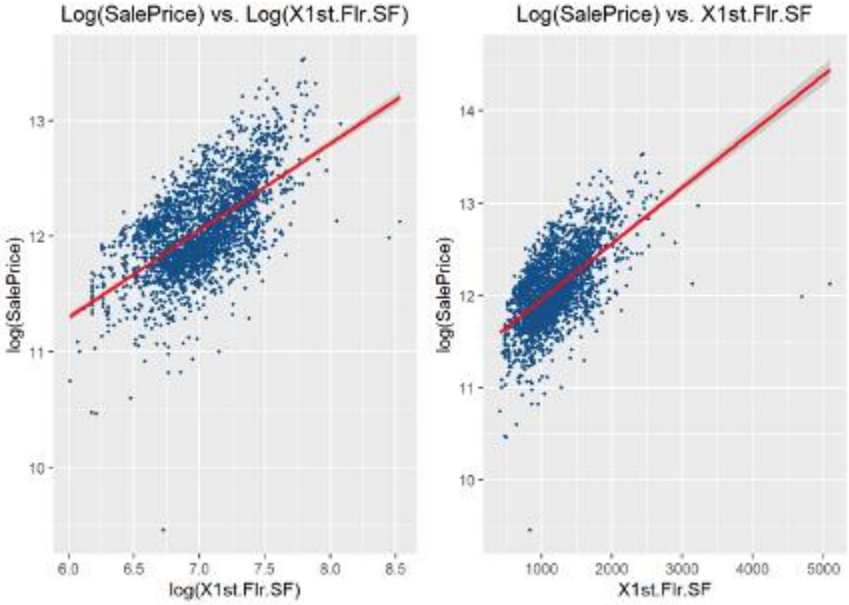


Figure 10. Comparison Of logarithmic transformation on Gr.Liv.Area


Log(SalePrice) vs. Log(Gr.Liv.Area)

Log(SalePrice) vs. Gr.Liv.Area

Figure 11. Comparison of Logarithmic Transformation of X1st.Flr.SF


Log(SalePrice) vs. Log(X1st.Flr.SF)

Log(SalePrice) vs. X1st.Flr.SF

Figure 12. Comparison of Logarithmic Transformation of Garage.Cars

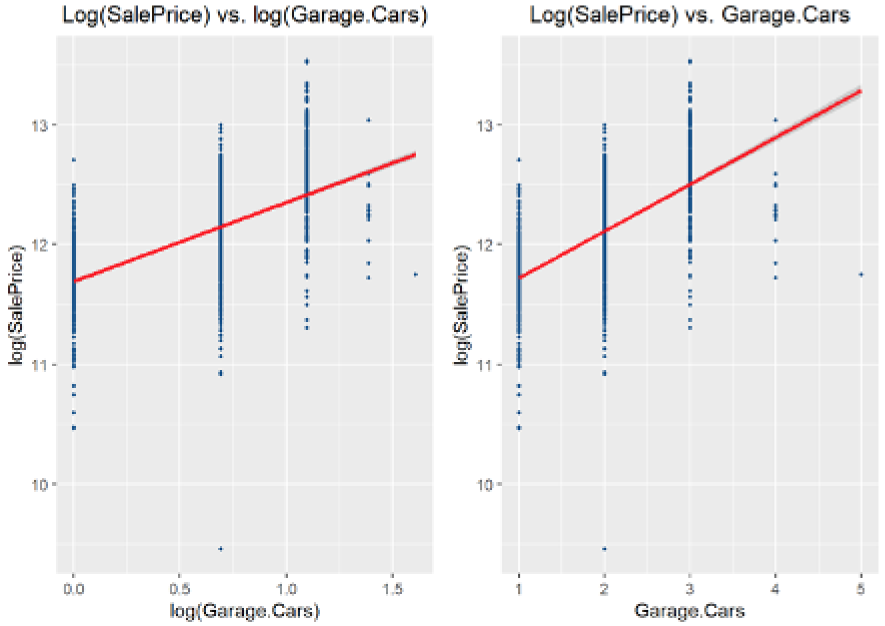


Figure 13. Comparison of Logarithmic Transformation of Garage.Area

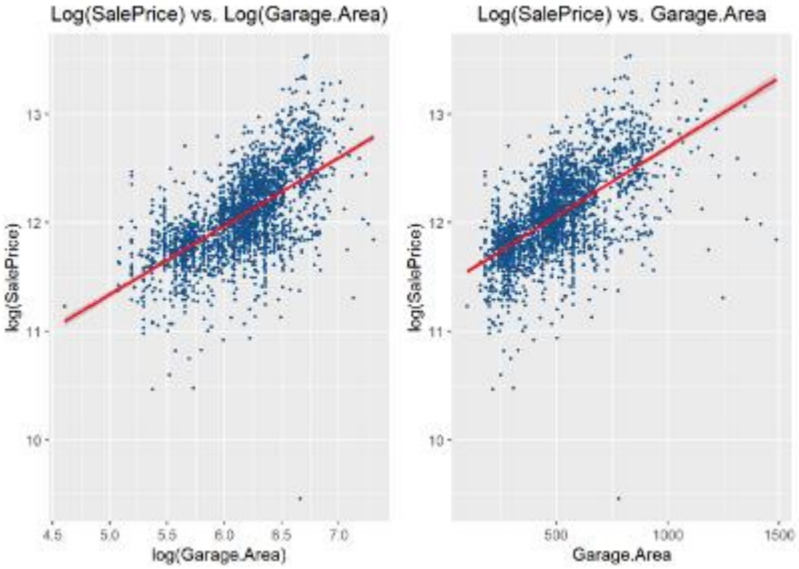
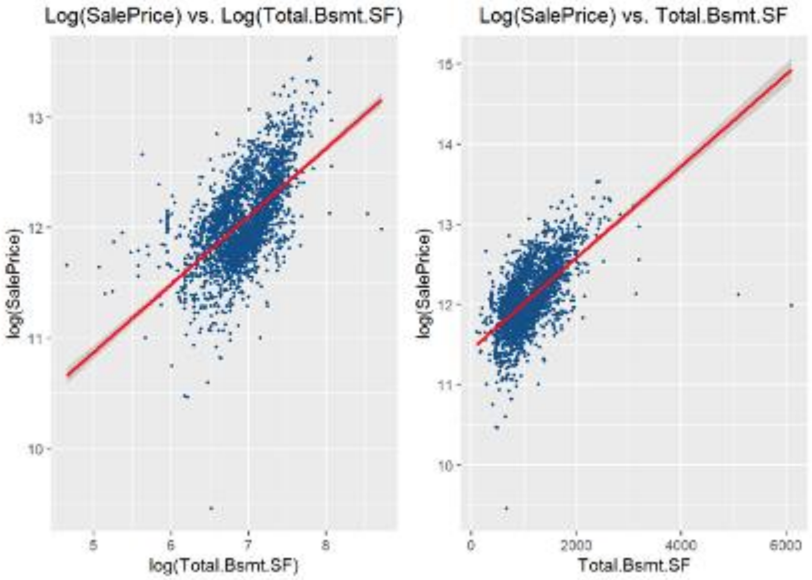


Figure 14. Comparison of Logarithmic Transformation of Total.Bsmt.SF



Appendix: Bibliography

Buuren & Groothuis-Oudshoorn “mice: Multivariate Imputation by Chained Equations in R” *Journal of Statistical Software,* vol 45, no 3, 2011.

DeCock, 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, doi:http://jse.amstat.org/v19n3/decock.pdf.

Ference, Audrey. "What Is a Comparative Market Analysis? The CMA Explained>." *Real Estate News and Advice | Realtor.com®*, Realtor.com, 11 Apr. 2019, www.realtor.com/advice/sell/understanding-the-comparative-market-analysis/.

McDonald, John F. "Market Value Websites: How Good Are They?" *Journal of Real Estate Literature* , vol. 14, no. 2, 2006, pp. 225–230., doi:http://jstor.org/stable/44103549.

Mehdi. "Ames Housing Prices." *Kaggle*, 6 Jan. 2018, www.kaggle.com/mnoori/ames-housing-prices.

Miller, Peter. "How Does a Real Estate Agent Set My Home Asking Price?" *Mortgage Rates, Mortgage News and Strategy : The Mortgage Reports*, 22 Sept. 2018, themortgagereports.com/42630/how-does-a-real-estate-agent-set-my-home-asking-price.