House Sales in Ames, Iowa

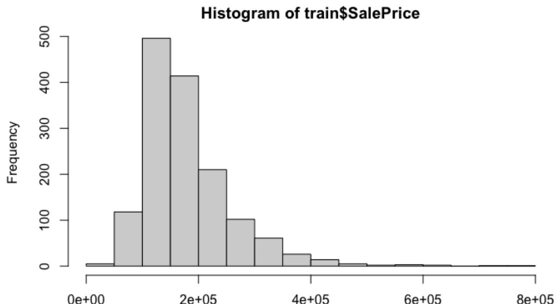
Andrew Scavo, Anthony Fini, John Mignone, Josh Cloute

ABSTRACT

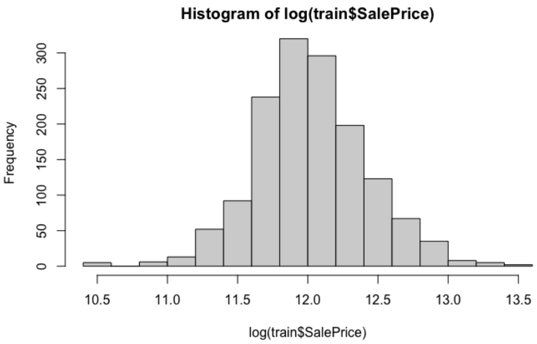
We ran a data set that contains multiple variables on the house sales of Ames, Iowa. What we were trying to find out is what determines the sales prices of these houses. Different factors that may go into these results are things such as; lot size, number of bedrooms, and square footage of the house. We used different analysis techniques to construct a model that would show the relationship with these factors on sales price. We use the R programming language to make a linear regression model to demonstrate these direct or indirect relationships. From our analysis we can better understand what homeowners value when they are shopping for a place to live. This can be beneficial to agents who are trying to sell homes, or even current homeowners who are trying to increase their home's resale value.

INTRODUCTION

To start, we chose a data set that we thought would work well in constructing our regression model. The data set comes from a 2011 study by Dean De Cock of Truman State University describing the sale of houses in Ames, Iowa from 2006 to 2010. This data set contains 2930 house sales and 80 explanatory variables, and of these variables there are

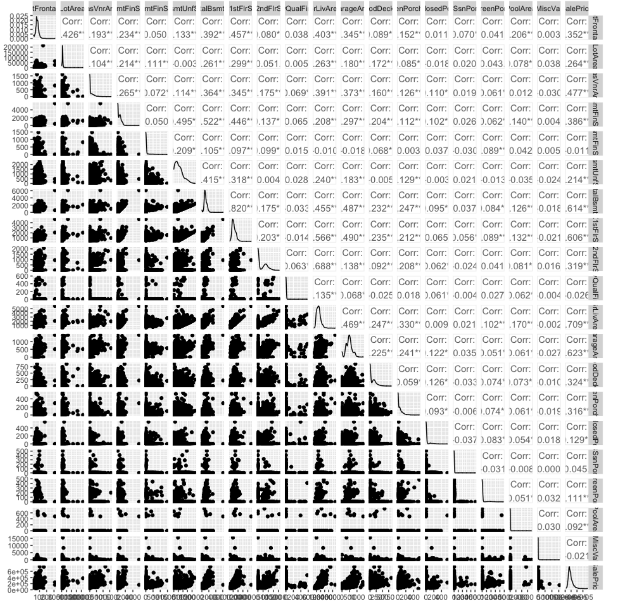


23 nominal, 23 ordinal, 14 discrete, and 20 continuous. This was a key reason why we chose this data set, as we have more than enough data to interpret. Our goal for the project was to construct a regression model to predict house sale prices in Ames, and we did this by running a series of regressions to make models that demonstrate the Sales Price. After doing this, we found the best possible model by conducting a series of tests comparing them with one another. Lastly, we used a 95% confidence interval with the model to include mostly accurate sales prices.



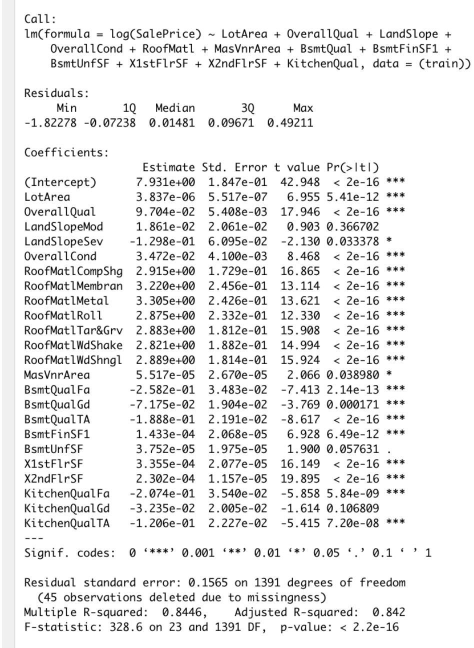
DISCUSSION

Firstly, there are some variables that we must include without question in our model, which mainly include variables that have to do with the size of the house (overall material and finish quality, total square feet of the basement area, above ground living area in square feet, total rooms above ground without including



bathrooms, and original construction date). As expected, all of these variables were significantly correlated with sales price, while not being highly correlated with each other. This is convenient because we do not want to include variables that are highly correlated with other variables in our model. For example, garage quality and garage cars were highly correlated with each other since a high quality garage will most likely be able to fit more cars, and since these variables are more correlated with themselves than sales price, so they will not be used. Additionally, when looking at other variables, some of the variable with the most positive correlation between themself and sales price are “GarageArea”, “1stFlrSf”, “2ndFl.Sf”, and “MasVnrType”. These variables also pertain to the size of the house, however they are also very correlated which presents an issue.

When starting the project, we noticed our response variable was not normally distributed, and because of this, we decided to transform the output by taking its logarithm. This did help create our response to look much more normally distributed than before. For analyzing the data set, we attempted to use many models, which include linear, logistic, validation, bootstrapping, and support vector machines. After running all these tests, we found using linear regression worked best when compared to all the other models. With the linear regression, we first ran the model with all of our variables to determine the most significant variables, and for this our R-squared was 0.95. Then, we ran the model again with the more significant variables, and the R-squared was 0.84. The significant variables referenced earlier include, but aren’t limited to LotArea, OverallQual, LandSlopeMod, OverallCond, RoofMatlCompShg, MasVnrArea, BsmtQualFa, BsmtFinSF, X1stFlrSF, and KitchenQualTA. After conducting this we used pcr regression with these significant variables.



```
Call:
lm(formula = log(SalePrice) ~ LotArea + OverallQual + LandSlope +
    OverallCond + RoofMatl + MasVnrArea + BsmtQual + BsmtFinSF1 +
    BsmtUnfSF + X1stFlrSF + X2ndFlrSF + KitchenQual, data = (train))

Residuals:
     Min       1Q   Median       3Q      Max
-1.82278 -0.07238  0.01481  0.09671  0.49211

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.931e+00  1.847e-01  42.948  < 2e-16 ***
LotArea          3.837e-06  5.517e-07   6.955 5.41e-12 ***
OverallQual      9.704e-02  5.408e-03  17.946  < 2e-16 ***
LandSlopeMod     1.861e-02  2.061e-02   0.903 0.366702
LandSlopeSev    -1.298e-01  6.095e-02  -2.130 0.033378 *
OverallCond      3.472e-02  4.100e-03   8.468  < 2e-16 ***
RoofMatlCompShg  2.915e+00  1.729e-01  16.865  < 2e-16 ***
RoofMatlMembran  3.220e+00  2.456e-01  13.114  < 2e-16 ***
RoofMatlMetal    3.305e+00  2.426e-01  13.621  < 2e-16 ***
RoofMatlRoll     2.875e+00  2.332e-01  12.330  < 2e-16 ***
RoofMatlTar&Grv  2.883e+00  1.812e-01  15.908  < 2e-16 ***
RoofMatlWdShake  2.821e+00  1.882e-01  14.994  < 2e-16 ***
RoofMatlWdShngl  2.889e+00  1.814e-01  15.924  < 2e-16 ***
MasVnrArea       5.517e-05  2.670e-05   2.066 0.038980 *
BsmtQualFa      -2.582e-01  3.483e-02  -7.413 2.14e-13 ***
BsmtQualGd      -7.175e-02  1.904e-02  -3.769 0.000171 ***
BsmtQualTA      -1.888e-01  2.191e-02  -8.617  < 2e-16 ***
BsmtFinSF1       1.433e-04  2.068e-05   6.928 6.49e-12 ***
BsmtUnfSF        3.752e-05  1.975e-05   1.900 0.057631 .
X1stFlrSF        3.355e-04  2.077e-05  16.149  < 2e-16 ***
X2ndFlrSF        2.302e-04  1.157e-05  19.895  < 2e-16 ***
KitchenQualFa   -2.074e-01  3.540e-02  -5.858 5.84e-09 ***
KitchenQualGd   -3.235e-02  2.005e-02  -1.614 0.106809
KitchenQualTA   -1.206e-01  2.227e-02  -5.415 7.20e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1565 on 1391 degrees of freedom
  (45 observations deleted due to missingness)
Multiple R-squared:  0.8446,	Adjusted R-squared:  0.842
F-statistic: 328.6 on 23 and 1391 DF,  p-value: < 2.2e-16
```

CONCLUSION

All in all, the main conclusion we made based on our models is that the main factor that correlates with the sales price of the house is the overall size, as many of our significant variables had to do with this. Because of this, we recommend adding any type of rooms or house extensions, as they will almost surely add as much value, if not more, to your home. Additionally, as another noteworthy finding, we found that houses are appreciating assets. Due to this, we recommend buying houses as soon as you can, as even if you increase your financial stability overtime, the house you were looking to purchase initially will only increase in price, thus having, at best, the same financial risk as before.