Zillow Project

Patrick Passafiume

November 1, 2017

Business Problem

A home value predictive model could have various applications in the real estate industry. Real estate agents could leverage that information in trying to buy/sell homes for clients. Home flippers can use that information when investing in homes and deciding what parts of the house should be flipped to maximize home value. Models predicting home values are used by sites like zillow and trulia to help customers in purchasing homes.

Zillow Data

The Zillow data that was pulled includes the zpid, address, state, city, zipcode, year built, the last sold price, tax assessment, lot size, finished square footage, bathrooms, bedrooms, and the zestimate. All data comes from comparable homes of a target address. Any NA values were replaced with the mean of that column.

Modeling Process

Initially, the values chose to be predictors included zipcode, year built, last sold price, tax assessment, lot size finished square footage, bathrooms, and bedrooms. Zestimate was used as the target variable. The first iteration of the model included all predictors without any transformations. The first model accounted had an R squared of 0.90. One zipcode and the bathrooms variables were not significant. Multicollinearity was tested to see if any predictors were highly correlated. Different methods were used to select the "best" predictors for the model. Since there weren't that many predictors, one method was trial and error to see what predictors were significant and how they affected the R squared value. Forward, backward, and stepwise AIC methods were also used to filter predictors. After experimenting with all the methods, the predictors that were used for the model were last price sold and finished square footage. This model accounted for 0.87 for the variation explained by the model. The model was tested by breaking the data into two sets, a training set and test set and compared to the actual zestimate value for each address.

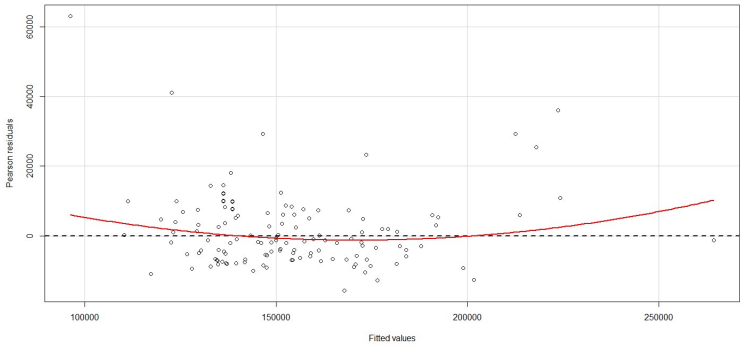
Model Interpretation

The coefficients for the model include last price sold with a value of 0.612 and finished square footage with a value of 16.45. For every 1 dollar increase in last sold price, increases the predicted home value by 61 cents. For every 1 square foot increase, the predicted home value increases by 16 dollars. Al of the predictor values are statistically significant.

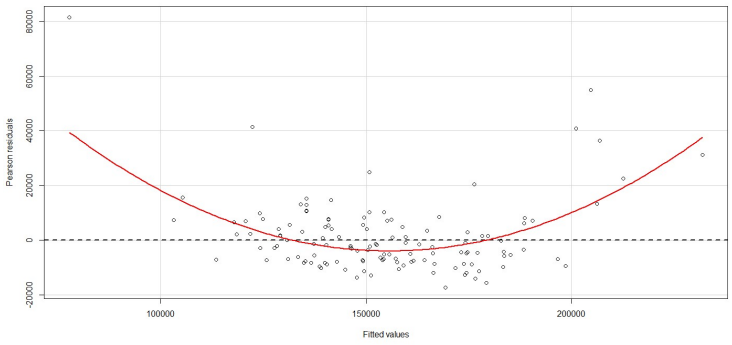
Model Evaluation

I feel like the model in not the most efficient or accurate. The predicted values did correlate very highly with the actual amounts, but that might be due to only having around 200 rows of data so overfitting could have happened. With more data, the model may be more accurate to predict home values. It breaks many of the regression assumptions. I tried to transform the data to make it more normal and built the model with the transformed data. After I looked at the residuals, it seemed like transforming the data made the model less accurate. Even after doing different transformation methods (log, square root, 1/data, etc), the model still did not meet the assumptions necessary for linear regression. There are many improvements that can be made to this model. One thing that should be done, is finding a better way to make the model fit the assumptions more effectively. There are potentially other variables that aren't in the Zillow data that could help make the model more accurate.

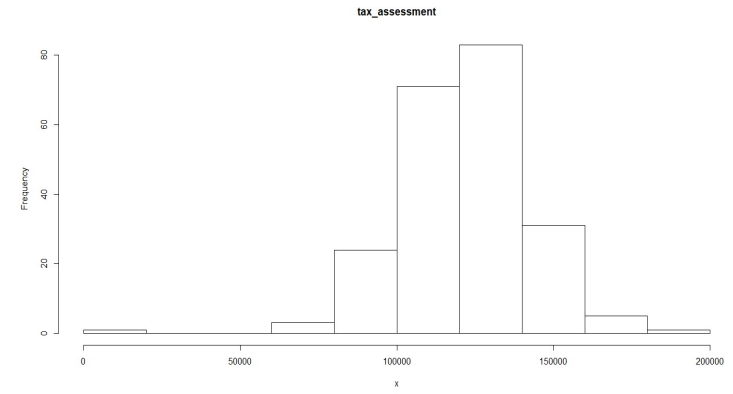
Residuals without transformations:

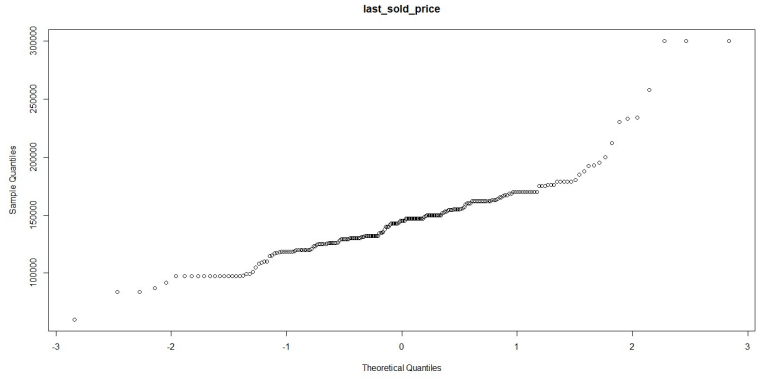


Residuals with transformations:

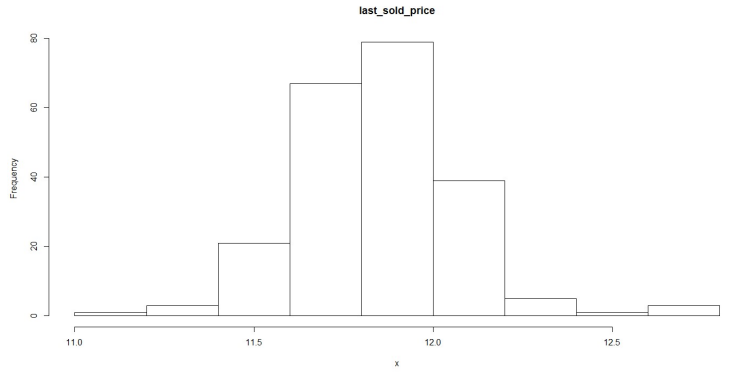


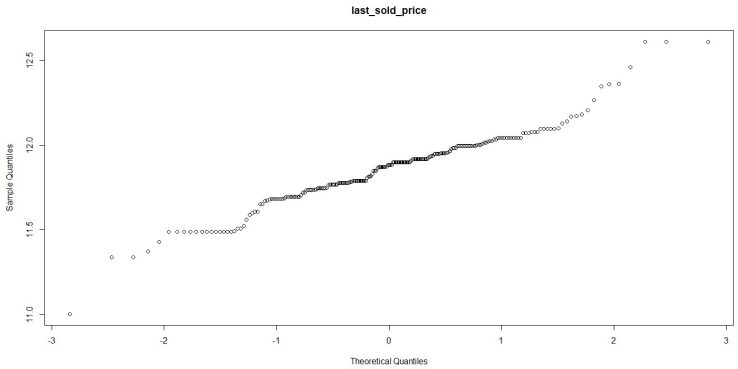
Last sold price before transformation



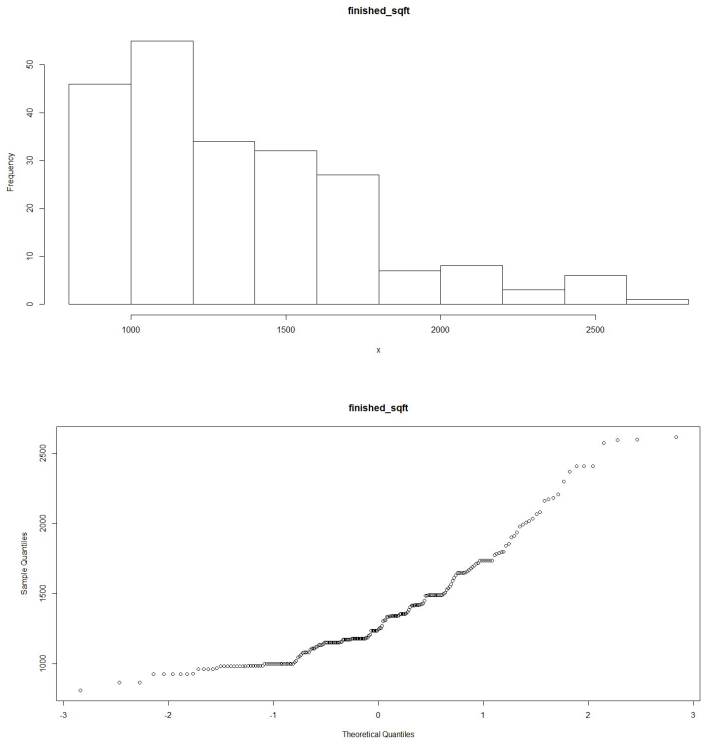


Last sold price after transformation

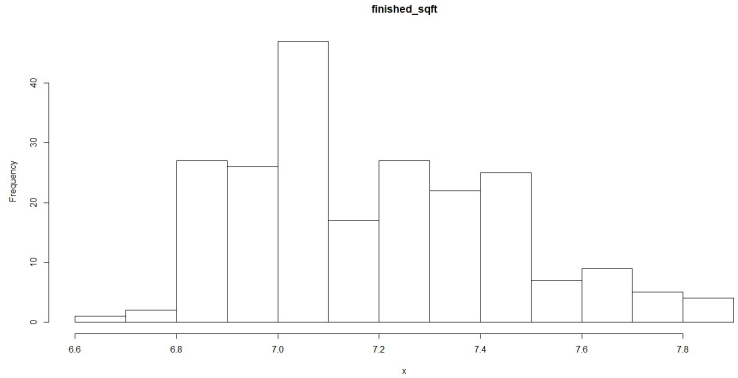


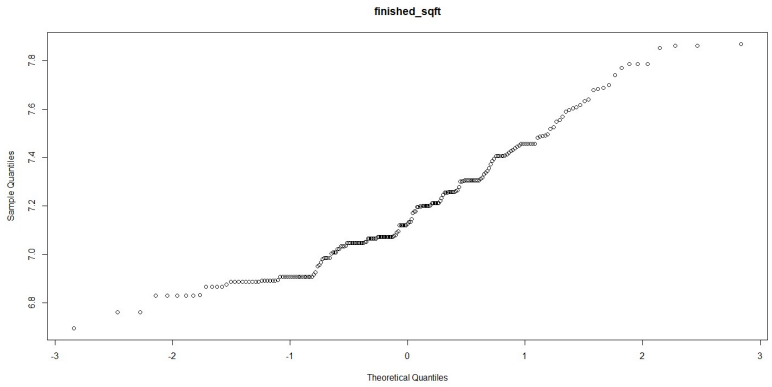


Finished square feet before transformation



Finished square feet after transformation




finished_sqft
Sample Quantiles
Theoretical Quantiles
6.8
7.0
7.2
7.4
7.6
7.8
-3
-2
-1
0
1
2
3