Development, Validation and Analysis of Linear Regression Model to Analyze Home Valuation in Northampton, MA

**BACKGROUND**

Modeling for home prices has often been sought after to gain insight in estimation of potential sale price for the buyer, seller, realtor, and those tied to the residential real estate industry. Through this analysis we will consider a sample set of home sales from Northampton, Massachusetts. Unique to this neighborhood is the Norwottuck Rail Trail. In 1997, the Rails to Trails Conservancy began to put more pressure on Massachusetts state to refurbish the worn-out railroad line as a "linear park." By 2007, the old Central Massachusetts Branch had become the Norwottuck Rail Trail and work is still being done to fully complete the 104-mile trail connecting Eastern and Western Massachusetts. Today the trail is enjoyed recreationally by many1.

Northampton consists of multiple zip codes. In this analysis, observations in zip codes 01060 and 01062 were represented within the data set. According to Google maps2 , the zip codes are about 3.7 miles away from each other, with vicinity to the trail system. Given this, we wish to investigate if distance to the Rail Trail can be shown to influence estimated home valuations within Northampton, Massachusetts. This will be achieved by development, validation, and analysis of a linear regression model with the available variables in the data set with any potential models to explicitly feature the distance variable.

**DATA**

The analysis used a data set of Northampton, Massachusetts home sales (value) provided in the “RailTrail” data file. The data set comprises of a total of 104 observations (n = 104) and 28 variables. The observations were from home sales in Northampton in the year 2007. The data is observational and assumed to be chosen at random from all sales within this neighborhood during 2007. Additionally, the data set consists of various variables describing the observations (homes) such as square feet, garage spaces, number of full and half baths, lot size, zip code, walk score, bike score, number of rooms, Zillow estimated prices in 1998, 2007, 2011, 2014, and distance to nearest point of entry to the rail trail network.

For this analysis the variable price 2014, Zillow Group’s estimated home valuation in 2014 dollars, was used as the response variable. Consideration for independent variable selection was mirrored by what Zillow Groupdeemed informative to its own neural network for modeling estimated home valuations. These home characteristics include square footage, location, or the number of bathrooms3. The independent variables available within the data set that mirror Zillow Group selections are distance, square footage, number of rooms, and number of bathrooms.

**PRELIMINARY MODEL INVESTIGATION**

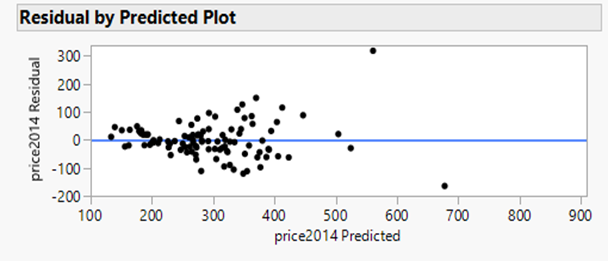
The preliminary model investigation we tested our variables selection of square footage, distance, number of rooms, number of full baths and number of half baths which is given by:

Model 1: E(y) = β0 + β1x1 + β2x2 + β3x3 + β4x4 + β5x5

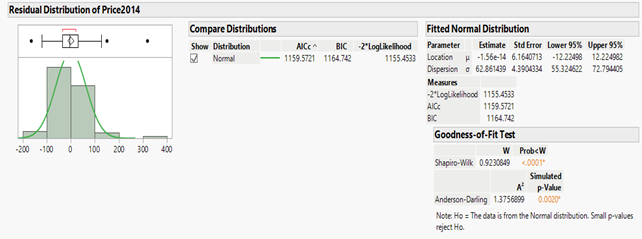
This yielded the following summary statistics as follows:



We can see from the model summary only square feet, distance and number of full baths were statistically significant in prediction of the observed response variable price 2014.



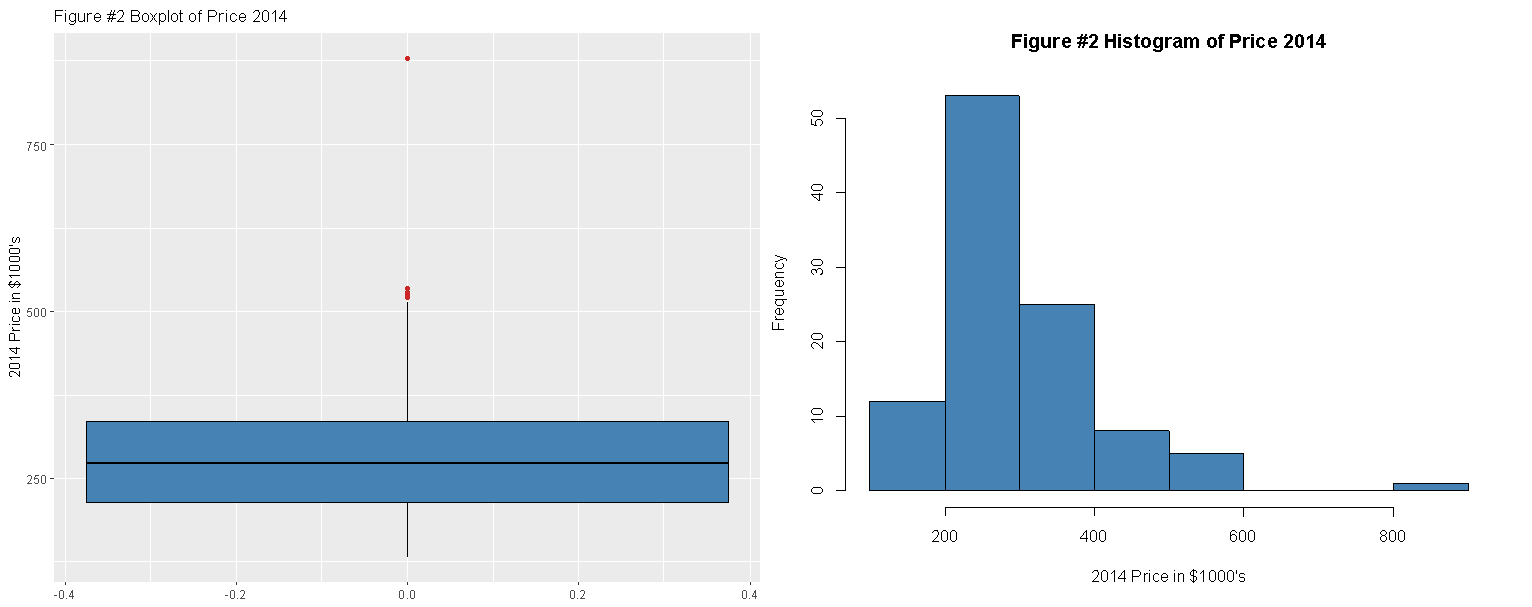
From the residual plot above, it is apparent that the linear regression assumption of homoscedasticity is violated. When there is unequal variance in the error terms at different levels of the predictor it introduces inconsistencies in the standard error & parameter estimates in the model. Subsequently, the confidence intervals and significance tests will be biased4.

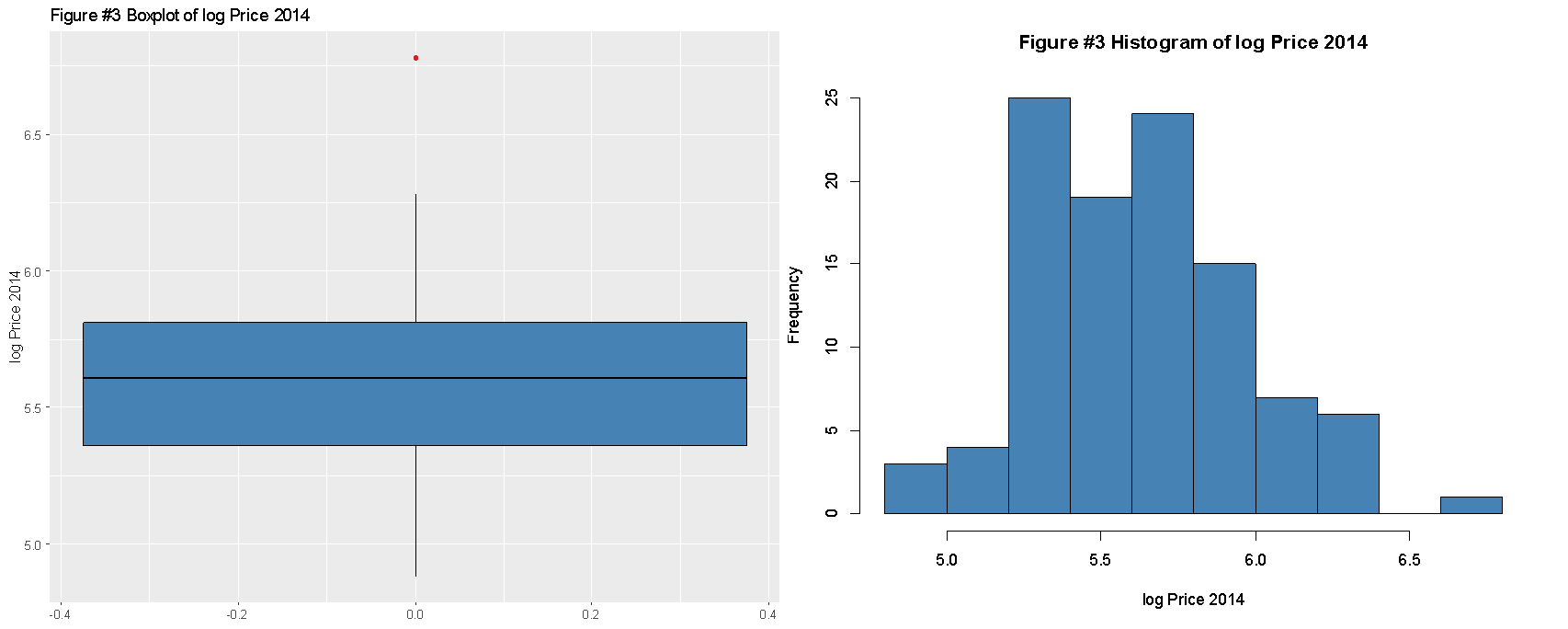


Furthermore, as shown by the distribution plot, the Shapiro-Wilk and Anderson-Darling test above, the assumption of residual normality is also violated. This makes it more difficult to determine whether the model coefficients are significantly different from zero4.

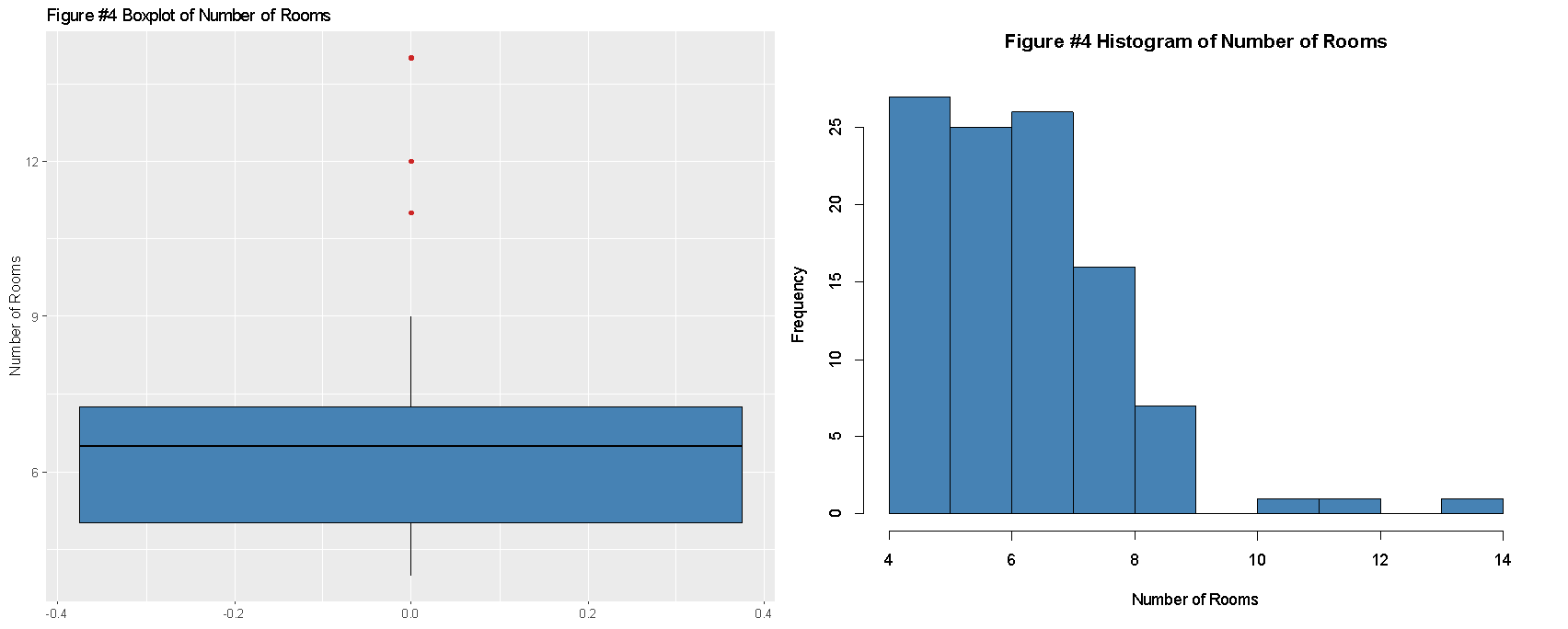
**VARIABLE TREATMENTS**

Because the preliminary model violated the regression assumptions of homoscedasticity and normality. The response variable price 2014 was transformed by applying the log function to the variable values. The log transformation reduces the impact of variability in the data and increases the normality of the response variability. The log transformation of price2014 also reduces the number of outliers (red points) in the variable from four to one (Figure 2,3).

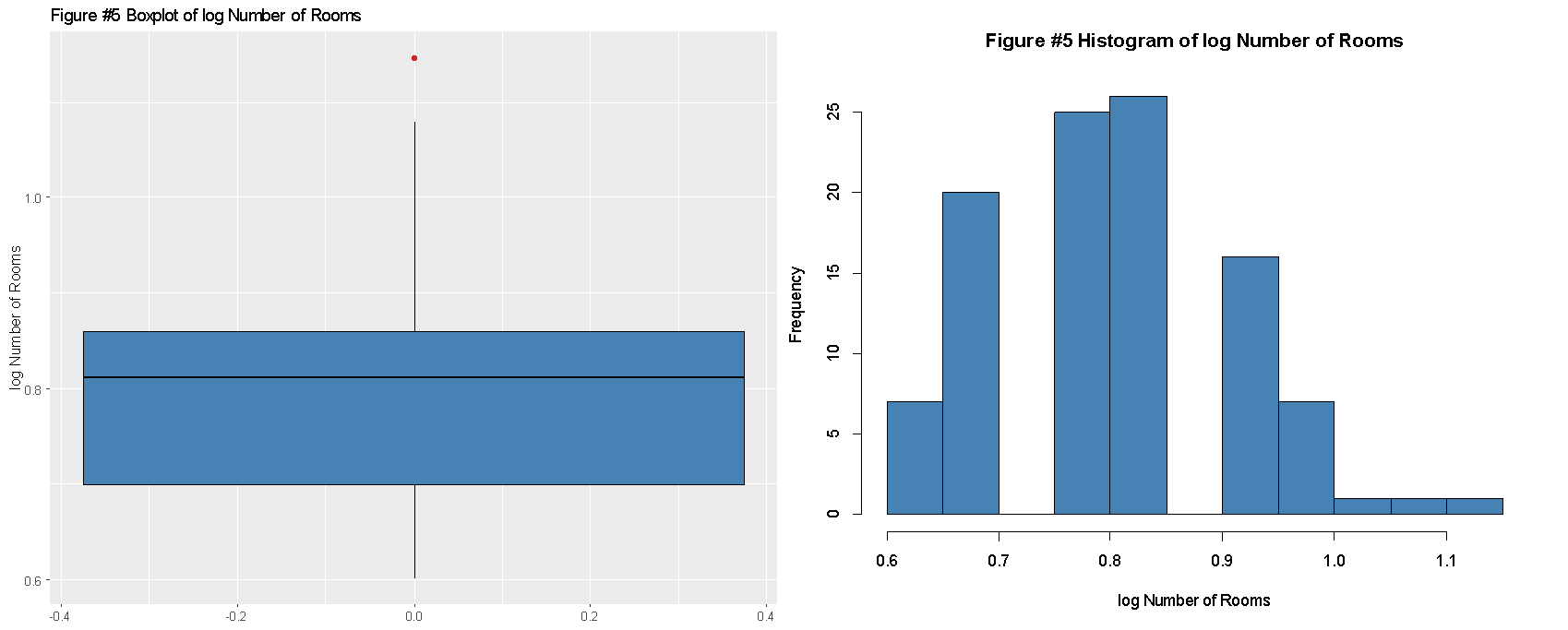




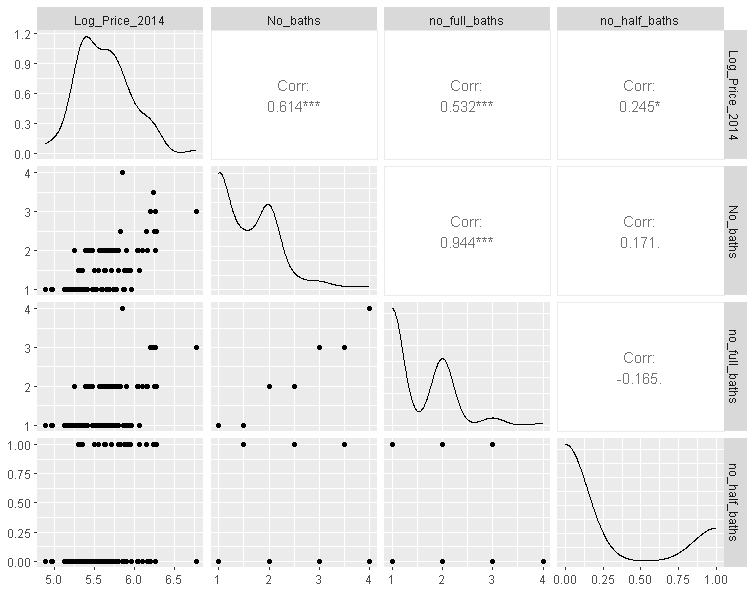
The same reasoning applies to the log transformation of the variable number of rooms. (Figure 4).



With the log transformation the number of outliers in the variable was reduced from three to one. (Figure 5).



The variables number of bathrooms and lot utilization are variables created from the available information in the “RailTrail” data set. It was found that combining number of full and half baths into one continuous variable had a stronger correlation metric with the response variable than number of full or half baths independently as is shown by the correlation matrix as follows:

Correlation Matrix.

Additionally, the variable number of bathrooms was created using the number full baths and number half baths by the following formula: number of bathrooms = full bathrooms + half bath\*0.5. Note that the variable half bath was coded as 1, has half bath or 0 if not in the dataset.

The variable lot utilization was created by using the data set variables acres and square feet by the following formula: lot utilization = (acres \* 43.56) / square feet (1000s). The constant term 43.56 is the total square footage within an acre scaled in thousands. This variable creation allowed for the information of house square footage and acres to be present in the model as these are common key features of a home and were found to be highly correlated to our response variable. With the addition of these two new variables, we selected the following variables for consideration in our second model, see Figure 6 below.

|  |  |
| --- | --- |
| Variable | Description |
| Y = log (price 2014) | Log of the Zillow estimated home price in 2014 |
| X1 = distance | distance (in miles) to the nearest entry point to the rail trail network |
| X2 = log (number of rooms) | Log of the number of rooms in a home |
| X3 = garage space | Number of garage spaces in a home |
| X4 = number of bathrooms | Total number of bathrooms in a home. Half bathrooms count as a half. |
| X5 = lot utilization | The ratio of total land size (square feet) to home size (square feet) |

Figure 6 Description of variables used in this analysis.

**MODEL REFINEMENT**

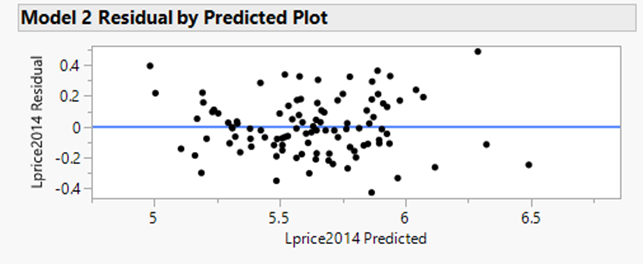
After applying the log function transformation to the variables price2014 and number of rooms. A second model was considered with the transformed variables and the newly created variables number of bathrooms and lot utilization. Which is given by:

Model 2: E(log(y)) = β0 + β1x1 + β2log(x2) + β3x3 + β4x4 + β5x5

This model yielded the following statistics:

We can see from the table above that all the predictors are found to be significant. The F-test of overall significance indicates that the model is statistically significant. The model can explain about 69% of the variation in the observed response variable. Further, the low standard errors in the β-estimates indicate that the model can more precisely estimate the unknown β value and indicates that the model fits the data well.

**MODEL VALIDATION**



From the residual plot above the heteroscedasticity of the residuals appears to have improved in comparison to model 1. A studentized Breusch-Pagan test for heteroscedasticity was performed with a significance level of = 0.05. With a p value of 0.12, we concluded that there is no evidence to reject the null hypothesis that the error variances are all equal.

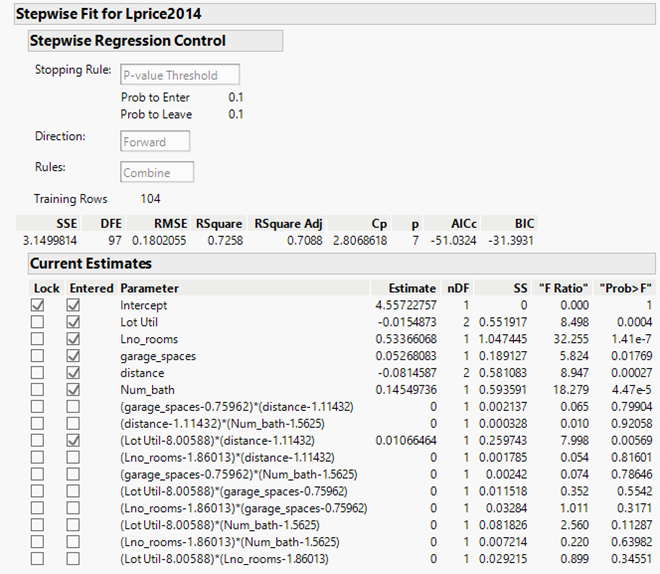
Graphical user interface, text, application

Description automatically generated

The residual distribution, as shown, that the Shapiro-Wilk, and Anderson Darling test there is enough evidence to conclude that the residual terms approximate a normal distribution.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model # | AICc | R2adj | PRESS | RMSE |
| 1 | 1170.62 | 0.66 | 0.61 | 64.45 |
| 2 | -45.14 | 0.69 | 0.66 | 0.19 |

The table above compares both models using metrics such as AICc, R2adj, PRESS, and RMSE. If we were to choose our model solely based on PRESS criteria, then model 1 would be slightly better than model 2. However, when we consider other criteria such as AICc, R2adj, and RMSE then model 2 outperforms model 1. Therefore, we will continue this analysis with model 2.



Using stepwise regression with a p-value of 0.1 to enter and exit. A significant interaction between lot utilization and distance was discovered.

Chart, line chart

Description automatically generated

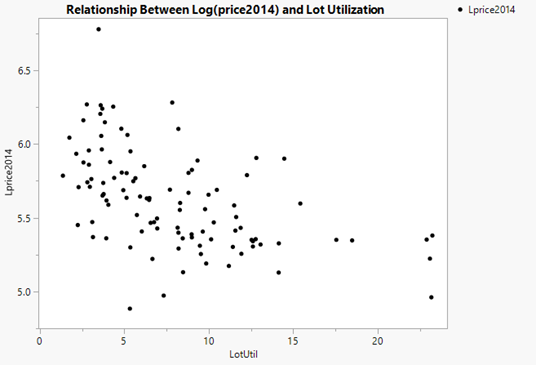
As observed in the interaction plot above, the general relationship between distance, lot utilization and the response variable appear to be negative. As distance increases and regardless of lot utilization, home prices decrease. As lot utilization increases the negative effects that these variables have on home price is less.

Chart, scatter chart

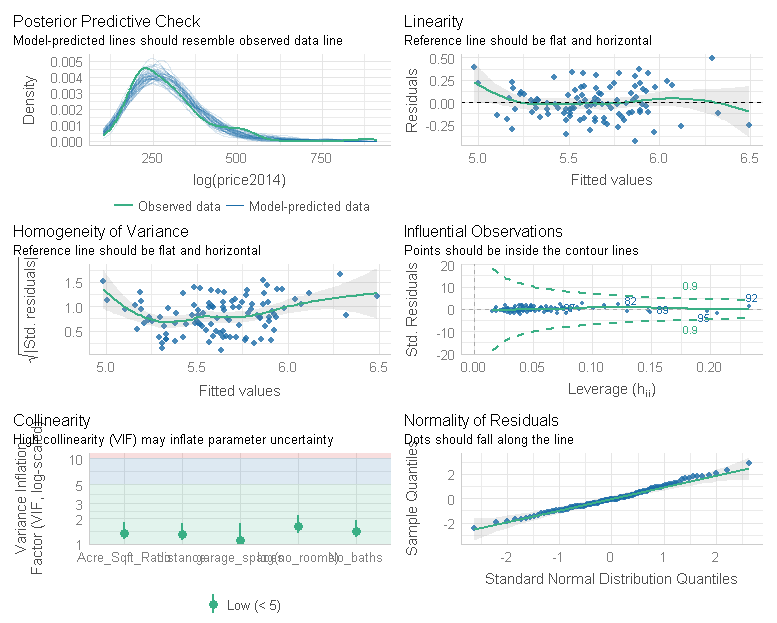
Description automatically generated

Lot utilization is defined as ratio of property acreage (in square feet) to house square footage. As the plot above shows, a higher lot utilization will usually indicate bigger property acreage or at least more undeveloped land. The interaction plot shows that properties with higher lot utilization generally have lower home prices. This seems counterintuitive as with bigger property you wouldn’t expect lower home prices.

The scatterplot below shows that the relationship between log(price2014) and lot utilization is possibly negative. It’s difficult to interpret because there are not many data points with lot utilization ratios greater than 12.5. To verify that the interaction between distance and lot utilization coincides with the data we’ll need more observations. Therefore, we elected to not include the interaction and use model 2 to determine the influence of distance on home price.

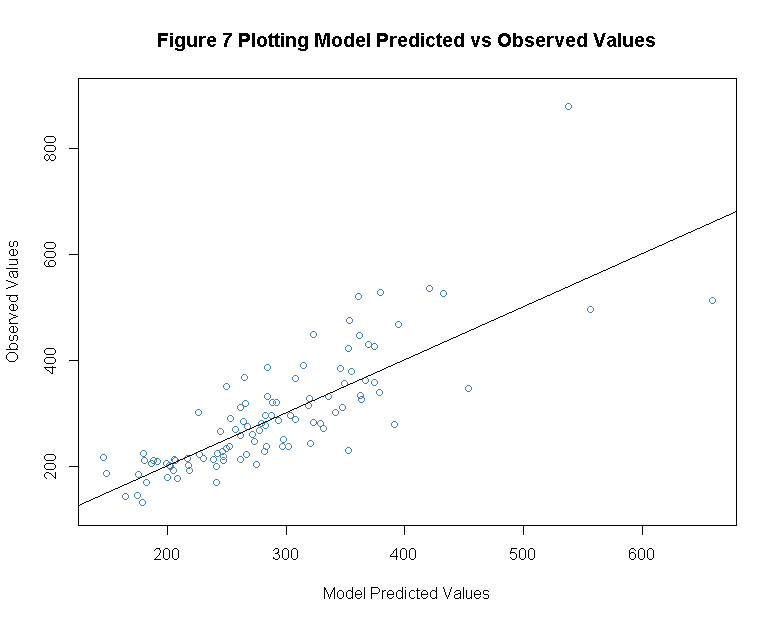


**MODEL VALIDATION**

Further assessment of model 2 utility, we check to see if any independent variables have collinearity to one another by means of variable inflation factor scores (VIF’s), Cook’s distance scores for influential observations, normality of the residuals via standard normal distribution quantiles as shown below.

The VIF scores are all below 2, indicating no high correlation between independent variables is present within the model. The Cook’s distance plot shows the residuals fall within the lines indicating there are no extremely influential data points that require further examination. The standard normal distribution quantiles plots show the residuals fall along the line. Therefore, we can infer the normal probability of the residuals suggest that the error term is normally distributed.

Plotting the model predicted values against the observed value price 2014 we see the model predicted values fall close to the regression line with exception of a few points towards the higher end of the line in Figure 6 below.



**MODEL COEFFICIENTS**

We can then interpret the beta coefficients as a percentage of change in estimated home valuation by e*b*-1 where all other variables are held constant.

* X1 distance we expect a decrease in home valuation of 6.53% for every one-unit increase in X1.
* X2 log number of rooms we expect an increase in home valuation of 25.24% for every 50% increase in number of rooms.
* X3 garage spaces we expect an increase in home valuation of 6.22% for every one-unit increase in X3.
* X4 number of baths we expect an increase in home valuation of 15.26% for every one-unit increase in X4.
* X5 lot utilization we expect a decrease of 1.26% for every one-unit increase in X5.

Lot utilization has the smallest on average impact on home valuation price where number of bathrooms has the highest on average effect on valuation. Distance seemed to be slightly better than the number of garage spaces, indicating distance is more of a possible consideration for those interested in homes within Northampton between the two. Lastly, the number of bathrooms and number of rooms has the highest effects on estimated home valuation in our model.

**CONCLUSION**

To conclude we have established statistically significant variables from our data set as well as creation of two new variables in which we found to be informative in the prediction of our response variable estimated home valuation price 2014. We were focused on if distance could be shown to have effect on the response variable, which our modeling process show could be done with significance. The estimated effect distance has on home valuation was shown to be 6.5% decrease for every one-unit increase in distance.

This information can aid in the real estate industry as this metric could be used as a first step screener of listings for clients interested in Northampton homes near the Rail Trail system. While this model proved to have utility in the future prediction of estimated house valuation, it is important to note that there could be other variables not within the data set that would prove to be informative, and likely influence the model’s performance metrics.

|  |  |  |
| --- | --- | --- |
| CITATIONS   1. Librarian, Visual Arts. “Research Guides: LSS240: 2018 Student Guide: Railroads and Rail Trails.” *Libguides.smith.edu*, libguides.smith.edu/lss240-spring2018/railroadstorailtrails. Accessed 19 Apr. 2023. 2. “Google Maps.” *Google Maps*, www.google.com/maps/dir/BAY+STATE+VILLAGE. Accessed 19 Apr. 2023. 3. “What Is a Zestimate? Zillow’s Zestimate Accuracy.” *Zillow*, [www.zillow.com/z/zestimate/](http://www.zillow.com/z/zestimate/). Accessed 19 Apr. 2023. 4. “Testing the Assumptions of Linear Regression.” *People.duke.edu*, people.duke.edu/~rnau/testing.htm. Accessed Apr. 2023. |  |  |