STATS 101A - Final Project Report: Predicting Austin House Prices

Audrey Lu, Bebel Yen, Joey Lee, Megha Ravi, Robert Alarcon, & Louisa Zhao

Introduction

The purpose of this project is to examine factors that significantly influence housing prices in Austin, TX, and create a model for housing price prediction. The Austin House Prices Dataset used for this analysis consists of real estate data, specifically focusing on single-family homes in Austin, TX. Our overarching research question is: what are the factors that significantly influence Austin house prices and what is the best model to predict them? The data includes key housing features such as the number of bathrooms, living area size (in square feet), average school rating, and property age. The dataset was scraped from Zillow in January 2021. Before model-fitting, the data was cleaned and filtered to remove inconsistencies and outliers.

We chose Linear Regression Modeling to understand the relationships between independent variables and house prices using untransformed variables (Model 1). We then applied Box-Cox and inverse response transformations (Model 2) to address model assumption violations in the first model. Finally, Model 3 was developed using forward selection with AIC and BIC criteria, introducing the age of the house as an additional predictor. This final model best adhered to statistical assumptions and improved model fit and interpretability.

Data Description

First, we examine the distributions of the response and predictor variables:

Variable	Mean	Standard Deviation
Latest Prices	511,520	425,611
Number of Bathrooms	2.71	1.01
Living Area (SqFt)	2,204	1,023
Average School Rating	5.78	1.86

Table 1. Variable means and standard deviations.

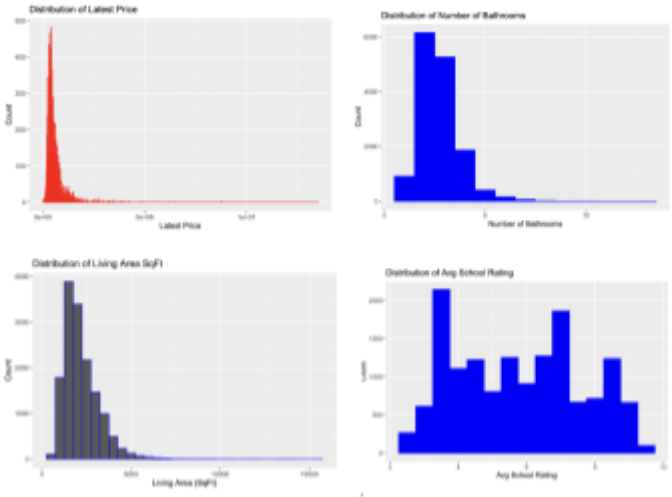

Figure 1: Distribution of variables

Let's examine the summary statistics. It seems that the response variable, house price, is still right skewed, while the number of bathrooms and living area is also generally right skewed. The average school ratings seem to have a multi-modal distribution. The summary statistics table summarizes each variable's mean and standard deviation. We can see that the average house price is about $511,520 , average number of bathrooms is 2.7, average living area is about 2204 square feet, and average school rating is about 5.7 out of 10. The standard deviations imply that most properties have between 1 and 4 bathrooms. There also seems to be some variation in school ratings, and significant variability in house size. Lastly, it also seems that there are houses whose prices are significantly lower or higher in comparison to the average.

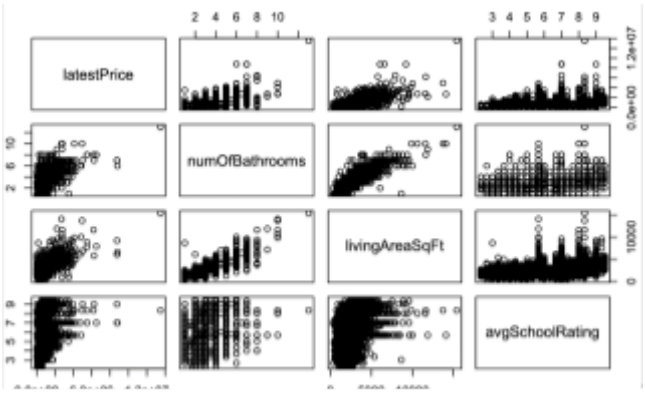

Figure 2. Scatterplot matrix of variables

From the scatterplot matrix, we can see a positive correlation between the number of bathrooms and the living area, as well as a weak correlation between the number of bathrooms and the latest price. There is a moderate positive correlation between living area and latest price. There is a weak correlation between the average school rating and the latest price. There does not appear to be a linear correlation between the number of bathrooms and average school rating, or living area square feet and average school rating.

Because the scatterplot matrix seems to suggest that the predictors have moderate positive correlations with housing prices, we chose to use multiple linear regression for our predictive model.

Results and Interpretation

For our first candidate model, we fit the untransformed response variable against the untransformed predictor variables:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.003e+05  1.021e+04  -9.826  < 2e-16 ***
numOfBathrooms   3.658e+04  4.866e+03   7.518 5.89e-14 ***
livingAreaSqFt   2.378e+02  5.147e+00  46.193  < 2e-16 ***
avgSchoolRating -1.953e+03  1.660e+03  -1.177    0.239
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 326100 on 14877 degrees of freedom
Multiple R-squared:  0.4123, Adjusted R-squared:  0.4121
F-statistic:  3478 on 3 and 14877 DF,  p-value: < 2.2e-16
```

Figure 3. Summary Statistics for Candidate Model 1

The corresponding linear equation is:

$$latestPrice = -100300 + 36580(numOfBathrooms) + 237.8(livingAreaSqFt) - 1953(avgSchoolRating)$$

The predictor variables numOfBathrooms, livingAreaSqFt, and avgSchoolRating are all significant with p-values < 0.05, while avgSchoolRating is insignificant with p-value = 0.239. **Overall F-test** is highly significant with p-value < 0.05, leading us to conclude that at least one predictor has a significant linear relationship with the response variable. R^2 of this model is 0.4123, indicating that it explains around 41.23% of variation in housing price.

We next examine diagnostic plots, outliers, leverage points, and influential points. The trend line in **Residuals vs Fitted Plot** appears mostly horizontal at 0, indicating linearity in predictor and response relationships. **Q-Q Plot** shows that standardized residuals follow a heavy-tailed distribution, suggesting normality of errors assumptions is violated. Heteroscedasticity can also be seen in the **Scale-Location** and **Residuals vs Fitted** plots.

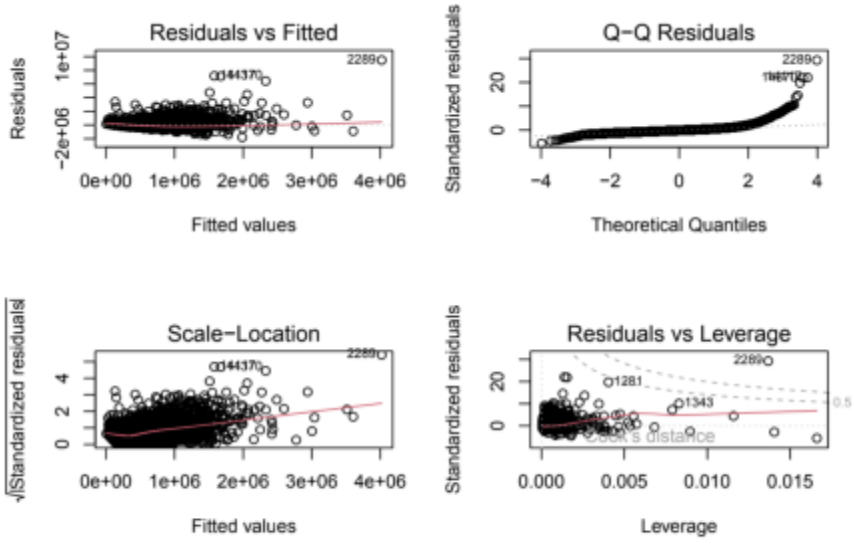

Figure 4. Diagnostic Plots of Candidate Model 1 (Untransformed)

Interestingly, Observation 2289 shows up on all four diagnostic plots. From the **Residuals vs Leverage Plot**, we can see that its Cook's Distance is greater than 1, indicating that it heavily influences the regression model due to high leverage and/or being an outlier.

Overall, this model has 428 outliers, 168 bad leverage points, and 555 influential points in all, which seems quite a lot. This indicates that many individual points do not fit the trend or have outsized influence on the regression.

We next used the Box-Cox method to see how variable transformation could help us fix this model's shortcomings.

```
bcPower Transformations to Multinormality
               Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
latestPrice      -0.0186       -0.02      -0.0313      -0.0060
livingAreaSqFt    0.1522        0.15       0.1292       0.1751
numOfBathrooms    0.4272        0.43       0.4012       0.4531
avgSchoolRating   0.6473        0.65       0.5976       0.6969
```

Figure 5. Power transformation statistics

Likelihood ratio test that all transformation parameters are equal to 0 has p-value <2.22e-16, indicating log-transformation of all variables is not appropriate. Likelihood ratio test that no transformations are needed for any parameters also has p-value <2.22e-16, indicating that some type of transformation is needed for at least one variable.

We decided to use the recommended rounded power transformations to fit our second candidate model, taking the log of the response variable since its rounded power is close to 0.

```
Call:
lm(formula = log(latestPrice) ~ numOfBathrooms_pt + livingAreaSqFt_pt +
    avgSchoolRating_pt)
```

All three transformed predictor variables have significant p-values. **Overall F-test** is also highly significant, so we reject the null hypothesis that no predictors are significant. However, R^2 has decreased slightly from our first candidate model to 0.3932, indicating that this model explains around 39.32% of the variation in latestPrice.

Observing the diagnostic plots (Figure B in Appendix), we found that this model violates many assumptions. So, we decided to implement inverse response plot to try a different response variable transformation.

We can see from the inverse response plot that a power transformation of 0.3 might be a little better than log for modeling latestPrice. We next fitted another version of Model 2, keeping the power transformations of predictors from Box-Cox, but transforming latestPrice by the power of 0.3 instead of log.

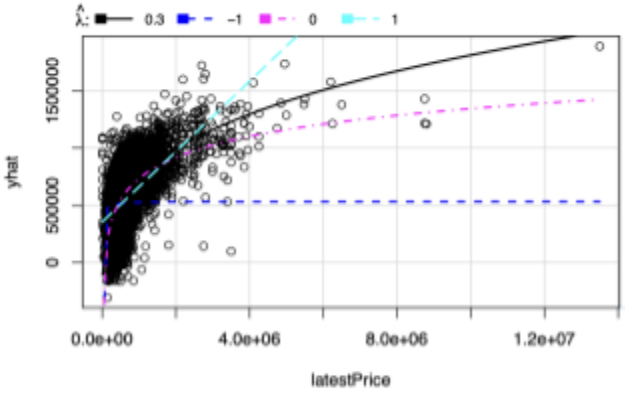

Figure 6. Inverse response plot (model 2, version 2)

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -22.70949    1.11531 -20.362   <2e-16 ***
numOfBathrooms_pt  4.10238    0.41101   9.981   <2e-16 ***
livingAreaSqFt_pt 19.10191    0.53373  35.790   <2e-16 ***
avgSchoolRating_pt  2.04103    0.09784  20.861   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.686 on 14877 degrees of freedom
Multiple R-squared:  0.402,  Adjusted R-squared:  0.4018
F-statistic:  3333 on 3 and 14877 DF,  p-value: < 2.2e-16
```

Figure 7. Summary statistics for model 2, version 2

We can see in the regression summary that all predictors as well as **overall F-test** are highly significant. R^2 shows the model explains around 40.2% of variation in latestPrice.

The final regression equation for Model 2 is:

$$latestPrice^{0.3} = -22.709 + 19.102(livingAreaSqFt)^{0.15} + 4.102(numOfBathrooms)^{0.43} + 2.04(avgSchoolRating)^{0.65}$$

Compared to the version of Model 2 with the log-transformed response, variance seems a little more constant in the **Scale-Location Plot**. In the **Q-Q Plot**, the magnitude of the deviation of the residuals has also reduced (compare Q-Q Residuals Plot in Figure 8 with Figure B in Appendix). Unlike with the log-transformed response, there aren't any standardized residuals with absolute values greater than 9, indicating the normality of errors is a little better. However, the heavy-tailed distribution suggests normality of errors assumption is still somewhat violated. **Residuals vs Leverage Plot** shows there are no points with very high Cook's Distance (over 0.5), which is a good sign.

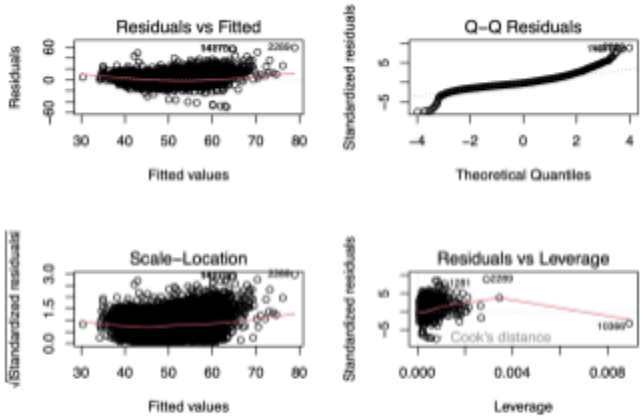

Figure 8. Diagnostic Plots of Candidate Model 2

Model 2 was an improvement over Model 1 in terms of adherence to model assumptions, despite the lower R^2. This is important since the reliability of inference tools derived from this model depends on satisfactory constant variance.

We next tried to see if bringing back variables we previously eliminated would help us improve our model. We also created a new variable, age, by subtracting the year the dataset was scraped, 2021, from the variable yearBuilt, giving the age in years of each house. We speculated there may be a preference for newer homes by homebuyers, which could possibly influence the relationship between the age and price of the home.

Using AIC and BIC as the criteria, we then used forward selection to build upon Model 2 to see which variables would contribute the most to our regression.

```
## Step:  AIC=54135.09
## latestPrice^0.3 ~ livingAreaSqFt_pt + numOfBathrooms_pt + avgSchoolRating_pt
##     age + propertyTaxRate + numOfStories + avgSchoolDistance +
##     lotSizeSqFt + numOfCommunityFeatures

## Step:  AIC=54138.11
## latestPrice^0.3 ~ livingAreaSqFt_pt + numOfBathrooms_pt + avgSchoolRating_pt +
##     age + propertyTaxRate + numOfStories + avgSchoolDistance +
##     lotSizeSqFt + numOfCommunityFeatures
```

Figure 8. Output for forward selection

Since both AIC and BIC decreased substantially after adding the age variable, we chose to fit our final model based on this suggestion.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -37.938793   1.111908  -34.12   <2e-16 ***
livingAreaSqFt_pt  21.138250   0.507582   41.65   <2e-16 ***
numOfBathrooms_pt   8.174437   0.401191   20.38   <2e-16 ***
avgSchoolRating_pt  1.679929   0.092755   18.11   <2e-16 ***
age                 0.117925   0.002758   42.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.308 on 14875 degrees of freedom
Multiple R-squared:  0.4677, Adjusted R-squared:  0.4676
F-statistic:  3268 on 4 and 14875 DF,  p-value: < 2.2e-16
```

Figure 9. Summary Statistics for Candidate Model 3

From the regression summary, all predictors are highly significant with p-values <2e-16. Overall F-test is also highly significant with p-value < 2.2e-16. R-squared is the best of all our models, indicating that 46.77% of variation in housing prices is explained by this model.

The regression equation for Model 3 is:

$$latestPrice^{0.3} = -37.939 + 21.138(livingAreaSqFt)^{0.15} + 8.174(numOfBathrooms)^{0.43} + 1.680(avgSchoolRating)^{0.65} + 0.118(age)$$

Residuals vs Fitted and **Scale-Location Plots** look pretty similar to Model 2, with residuals more evenly and randomly scattered than Model 1, but not to the point where we would conclude that variance is sufficiently constant. **Q-Q Plot** is also similar to Model 2, showing a slightly heavy-tailed distribution of residuals, indicating that normality of errors assumption is still not met. Since the trend line in the **Residuals vs Fitted Plot** is quite straight and horizontal at 0, we conclude linearity assumption is met.

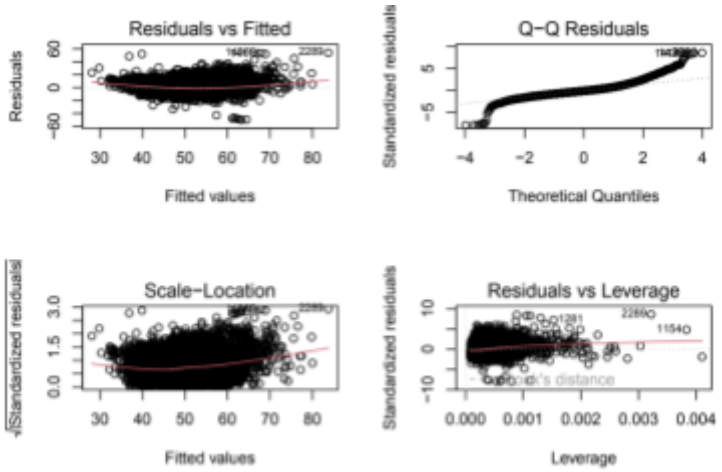

Figure 10. Diagnostic Plots for Candidate Model 3

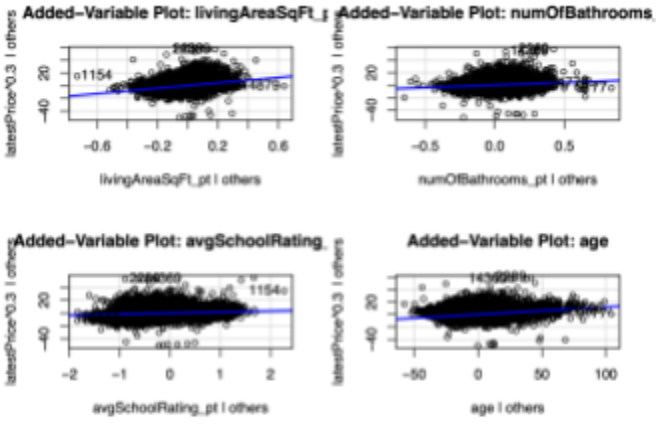

Figure 10. Added variable plots for Candidate Model 3

VIFs below 5 imply that multicollinearity among the predictors is not a big issue. However, observing the Added-Variable Plots, the subtle slopes on numOfBathrooms, avgSchoolRating, and age imply that they do not contribute as much to the regression as livingAreaSqFt, after adjusting for other predictors' effects.

Finally, we compared adjusted R^2, AIC, AICc, and BIC to decide which candidate model was the best. From the four criteria, we can see clearly that model 3 is the winner with highest adjusted R^2, lowest AIC, AICc, and BIC.

	AdjR2	AIC	AICc	BIC
1	0.4121341	377830.55	377830.55	377860.98
2	0.4018403	56554.32	56554.32	56584.75
3	0.4675778	54818.36	54818.36	54856.40

Figure 11. Table of model fit statistics across candidates

Therefore our final model is:

$$latestPrice^{0.3} = -37.939 + 21.138(livingAreaSqFt)^{0.15} + 8.174(numOfBathrooms)^{0.43} + 1.680(avgSchoolRating)^{0.65} + 0.118(age)$$

Discussion

Due to the power transformations in the final model, the coefficients are hard to interpret. Our attempt at interpretation is: a 1% increase in living area results in approximately (21.138 * 0.15%) = 3.171% increase in the 0.3 power of housing price, a 1% increase in number of bathrooms leads to a (8.174 * 0.43) = 3.15% increase in the 0.3 power of housing price, a 1% increase in average school rating leads to a (1.68 * 0.65) = 1.09% increase in the 0.3 power of housing price, and each additional year since the home was built adds 0.118% increase in the 0.3 power of housing price.

The positive correlation of living area size and housing price aligns with the real world, as larger homes tend to be more expensive. The positive correlation between high school ratings and increased prices also makes sense because better education quality increases real estate valuation. In fact, a 2005 study shows how environmental factors, including school quality, influence housing prices (Brasington & Hite, 2005). House ages have a small but positive effect, which is somewhat counterintuitive and may indicate that older homes that are historic or in desirable areas maintain their value or appreciate over time. A research study by Coulson and Lahr explores how historic designation affects property values, which shows how older homes in well-established neighborhoods retain value due to factors like mature landscaping (Coulson & Lahr, 2005).

The limitations of this model include the fact that there are many outliers and influential points in the final model. While extreme values were removed, housing price data often includes anomalies (e.g., luxury homes), which may skew predictions. Also, the predictor variables were fairly limited, as the model does not include possibly relevant factors such as crime rates, proximity to public transport, economic trends, etc., which can significantly impact home values. Future work could explore geospatial modeling, incorporating neighborhood characteristics and economic indicators to increase the accuracy of the model and prediction reliability.

Appendix: Additional Plots

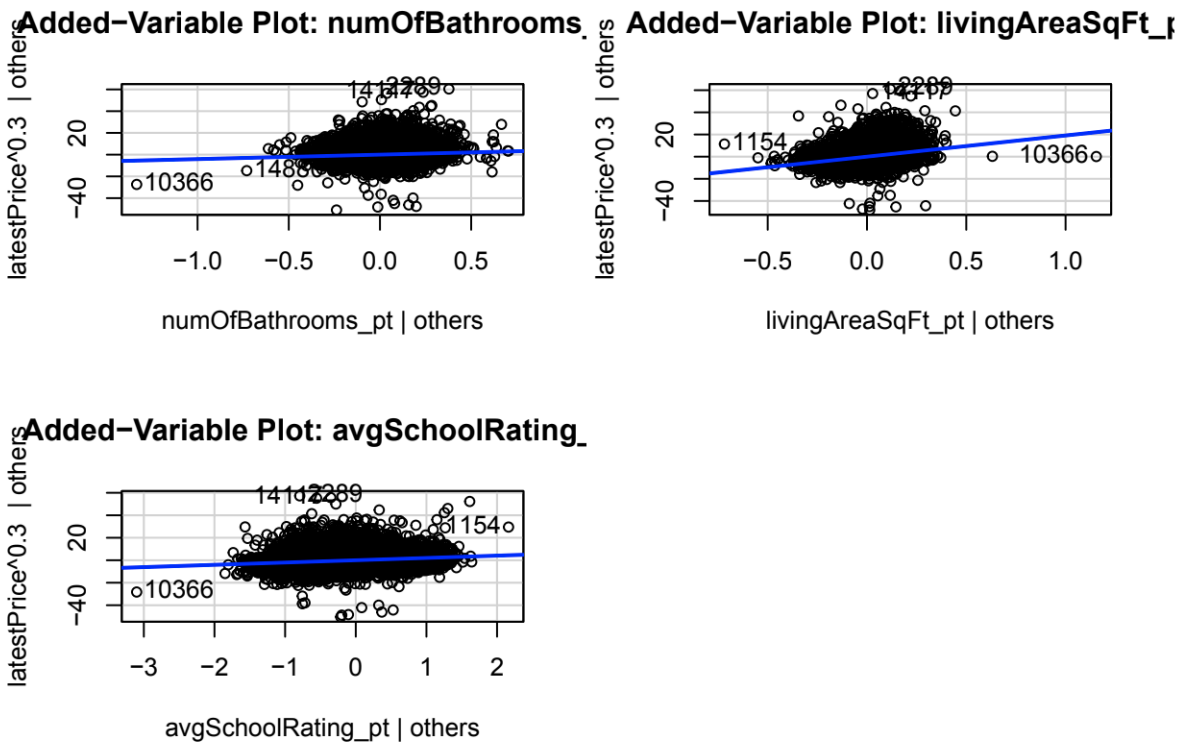

Figure A. Added Variable Plots for Candidate Model 2

The Added-Variable Plots show that the direction of the slopes are consistent with their corresponding regression coefficients, a sign that there is likely not heavy multicollinearity in our model. VIFs are also all below 5, indicating that multicollinearity is unlikely to be an issue. However, numOfBathrooms and avgSchoolRating have very weak slopes, suggesting that these variables do not seem to contribute as much to the regression as livingAreaSqFt.

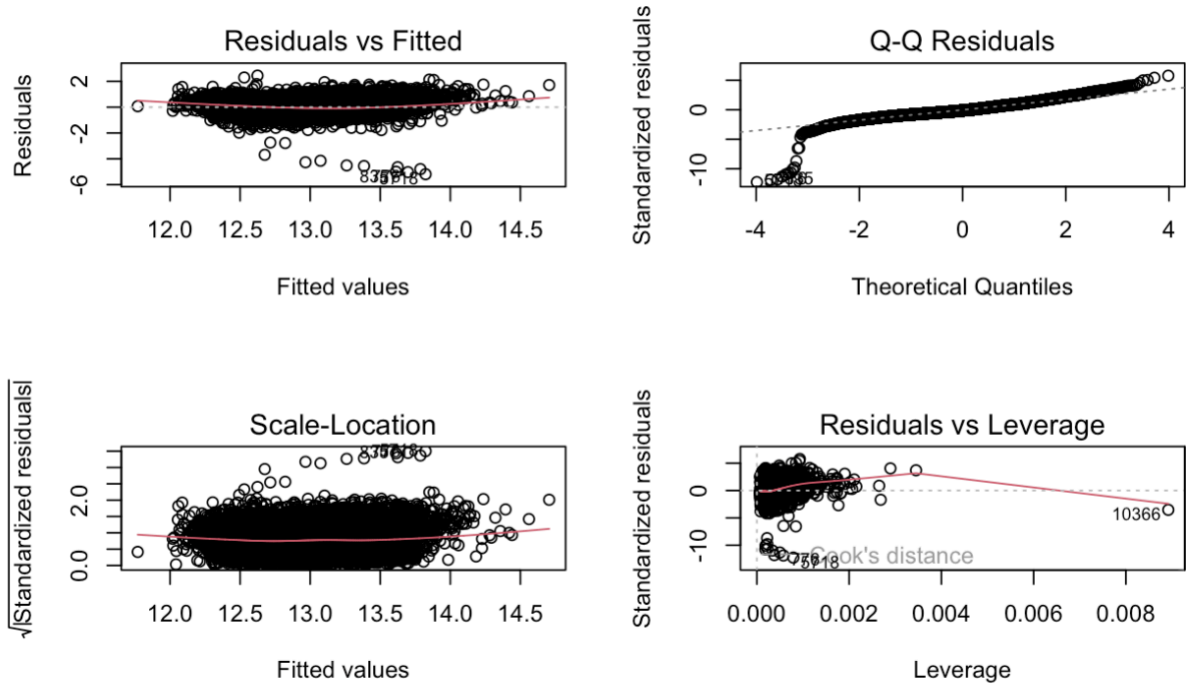

Figure B. Diagnostic Plots for Candidate Model 2, Version 1

With the first version of Model 2 fitted with a log-transformed response variable, we noticed violation of model assumptions especially regarding normality and constant variance. This prompted us to use inverse response plot to try a different transformation of the response variable.