STATS 6371 Final Project Report

Kyle Evans and Eric Graham

2024-12-09

Introduction

Given a dataset of home prices and features in Ames, Iowa, we set out to analyze existing factors impacting home prices and look for predictive models that can accurately predict home prices. Our first analysis (Analysis 1) is descriptive in nature: specifically, we are interested in understanding the effect that living area has on home prices in three neighborhoods in Ames. Our second analysis (Analysis 2) is predictive: we explored several linear regression models of varying size and complexity in an effort to find the best model for predicting home prices.

Note on Code Appendix and Figures

The assignment calls for an appendix to include our code, but because the code for our project is quite long, we have instead opted to include the entire notebook on our project github, which can be found here.

In order to keep the written paper at seven pages, we have moved our figures to the appendix to this document.

Data Description

Our dataset includes 1,460 rows and 81 features, 43 of which are characters (categorical variables) and 38 of which are numeric variables. In addition to including the sales price for each observation, the 80 other features cover a variety of aspects of the home, including the neighborhood, living area (in square footage), the number of bedrooms, the size of the garage, the lot frontage, and many others. This is a dataset that is very popular for predictive modeling and has been used in many Kaggle competitions.

To that end, we were also provided with a "test" dataset of 1459 observations which omits the sales price; we used this test data to validate our predictive models from Analysis 2 on Kaggle, and will include our Kaggle scores in evaluating those models.

Analysis 1: Descriptive Analysis

Introduction

Our first analysis examines the relationship between LogSalePrice and GrLivArea (measured in increments of 100 square feet) for homes in the NAmes, Edwards, and BrkSide neighborhoods. Specifically, the goal is to determine whether the relationship varies by neighborhood and to provide estimates with confidence intervals.

Transformation of SalePrice

An initial look at the data shows that the relationship between SalePrice and GrLivArea is positive and linear for all neighborhoods, with some outliers. However, because the distribution of SalePrice was right-skewed, we decided to use a log transformation of SalePrice in our analysis. See Figure 1 and 2in the appendix.

Outliers in Edwards Neighborhood

We identified two outliers (observations 524 and 1299) in the Edwards neighborhood for which an exceptionally large living area was associated with a lower-than-expected sale price. Extensive analysis of these influential points can be found in our code appendix. See figure 3 in the appendix.

We removed both of these outliers from the dataset. For the sake of presentation to the client, we would disclose these outliers, and note that the model is based on a dataset that doesn't include them. By removing these two data points, we increased the adjusted R2 of our model by 8%. In keeping with the idea that "all models are wrong but some are useful", we believe that the best course of action is to present this model to the company with the caveat that it should only be used with houses 3000 square feet or less. See figure 4 in the appendix.

General Model

We fit the general model as follows:

$$\log(\text{SalePrice}) = \beta_0 + \beta_1 \times \text{GrLivArea}_{\text{BrkSide}} + \beta_2 \times \text{GrLivArea}_{\text{Edwards}} + \beta_3 \times \text{GrLivArea}_{\text{NAmes}} + \epsilon$$

Results

The results of our model are shown in Figure 5 in the appendix, as well as the code in our Github.

The results show that the relationship between GrLivArea and LogSalePrice is positive for all neighborhoods, and the p-values of all coefficients is statistically signifcant. However, the magnitude of the effect varies by neighborhood: the effect of GrLivArea on LogSalePrice is smallest in the NAmes neighborhood and largest in the Edwards neighborhood. Our neighboorhood-specific models are as follows:

BrkSide

$$\log(\text{SalePrice}) = \beta_0 + \beta_1 \times \text{GrLivArea}$$

$$\log(\text{SalePrice}) = 10.79 + 0.000738 \times \text{GrLivArea}$$

Edwards

$$\log(\text{SalePrice}) = (\beta_0 + \beta_{\text{Edwards}}) + \beta_2 \times \text{GrLivArea}$$

$$\log(\text{SalePrice}) = (10.79 + 0.2339) + 0.0005387 \times \text{GrLivArea}$$

NAmes

$$\log(\text{SalePrice}) = (\beta_0 + \beta_{\text{NAmes}}) + \beta_3 \times \text{GrLivArea}$$

$$\log(\text{SalePrice}) = (10.79 + 0.6517) + 0.0003241 \times \text{GrLivArea}$$

Assumptions

Linearity

The plot of residuals against fitted values hows no apparent pattern, which suggests that the assumption of linearity is met (see Figure 6 in the appendix).

Normality

The Q-Q plot indicates that residuals closely follow a normal distribution, with some deviation at the extremes (see Figure 7 in the appendix).

Variance

The spread of residuals in the residuals versus fitted plot (Figure 6) is relatively consistent across fitted values, which suggests that the assumption of constant variance is met.

Independence

We will assume that observations are independent enough to meet this assumption, however there is a possible cluster effect due to houses being in the same neighborhood. By adding the Neighborhood variable to the model, perhaps the violation, if present, is somewhat mitigated.

Influential Points Analysis

After removing the significant outliers mentioned above (observations 524 and 1299), the Cook's D and outlier-leverage diagnostics reveal a few remaining influential points. However, these points were less impactful than those already removed, and while they are outside the normal range, we didn't deem them extreme enough to warrant removal (see Figure 7 and 8 in the appendix).

Parameter Estimates and Model Interpretation

The model estimates suggest significant relationships between log-transformed sales price and both neighborhood and living area. The intercept 10.79 represents the average log-sale price for homes in the BrkSide neighborhood (our reference category) when the living area is zero. This is obviously not practically feasible, and it exists solely as a parameter in the model.

The coefficients for the Edwards (0.2339) and NAmes (0.6517) neighborhoods indicate higher baseline log-sale prices compared to BrkSide. Additionally, the interaction terms for living area show positive effects on log-sale price, with the strongest effect in BrkSide (0.000738) and decreasing effects in Edwards (0.000539) and NAmes (0.000324).

Our 95% confidence intervals for the coefficients are as follows (meaning that we are 95% certain that the true coefficient is within the given range):

Term	Confidence Interval (95%)
Intercept	10.63 and 10.95.
NeighborhoodEdwards	0.0210 and 0.4468.
NeighborhoodNAmes	0.4708 and 0.8327.
NeighborhoodBrkSide:GrLivArea	0.000611 and 0.000866.
NeighborhoodEdwards:GrLivArea	0.000432 and 0.000645.
NeighborhoodNAmes:GrLivArea	0.000264 and 0.000384.

Conclusion

It is estimated that for every 100 square feet increase in living area, the median sales price of a home in Brookside increases by 7.38% (p-value < 0.0001). A 95% confidence interval for the true median increase in sales price is between 6.11% and 8.66%.

It is estimated that for every 1100 square feet increase in living area, the median sales price of a home in Northwest Ames increases by 3.24% (p-value < 0.0001). A 95% confidence interval for the true median increase in sales price is between 2.64% and 3.84%.

It is estimated that for every 100 square feet increase in living area, the median sales price of a home in Edwards increases by 5.39% percent (p-value < 0.0001). A 95% confidence interval for the true median increase in sales price is between 4.33% and 6.45%.

Analysis 2: Predictive Analysis

We now turn to the predictive analysis of home prices in Ames, Iowa. We will explore several linear regression models of varying size and complexity in an effort to find the best model for predicting home prices. The below table is contains the metrics for the models we tested and what follows after is a brief discussion of each model:

Linear Model Comparison Table

Note, that metrics were pulled from the resulting CARET models after LOOVC, not directly from the LM() function. For consistency, unless otherwise noted, all models were ran without the same outliers found above, 1299 and 524. The names in the tables should match up nicely in the outline of the project.rmd.

Model	Adjusted R^2	AIC	PRESS	Kaggle Score
SLR-M1: GarageArea	0.4301	647.7850	133.0714	0.31527
MLR:M0: GrLivArea + FullBath	0.5551	287.6378	103.9207	0.29097
MLR:M1: 50+	0.937	-2406.90	19.74	0.14395
MLR:				
MLR:				

Model 1: Simple Linear Regression

Our simplest model is a linear regression of SalePrice on GarageArea. This model seeks to predict home prices based on the size of the garage.

General Model

We fit the general model as follows:

$$\log(\text{SalePrice}) = \beta_0 + \beta_1 \times \text{GarageArea} + \epsilon$$

Results

As shown in the code in our appendix, we implemented Leave One Out Cross-Validation on our predictive models to evaluate their performance. The results of our model are also shown in Figure 9 in the appendix.

The results show that the relationship between GarageArea and LogSalePrice is positive, and the p-values for both the coefficient and intercept are statistically significant. The model has an adjusted R2 of 0.4301, which means that 43% of the variance in log-transformed sale price can be explained by the size of the garage. This is less promising than our descriptive model, but still a good start.

The linear regression of our equation is as follows:

$$\log(\text{SalePrice}) = 11.44 + 0.001236 \times \text{GarageArea}$$

Assumptions and Influential Points Analysis

Linearity The plot of residuals against fitted values shows no apparent pattern, indicating that the assumption of linearity is reasonably satisfied (see Figure 10 in the appendix).

Normality The Q-Q plot suggests that residuals follow a normal distribution, however some deviation is observed at the tails (See Figure 11 in the appendix).

Variance The spread of residuals in the residuals plot (shown in Figure 10) is relatively uniform across fitted values, indicating that the assumption of constant variance (homoscedasticity) is met.

Independence We will assume that observations are independent enough to meet this assumption, however there is a possible cluster effect due to houses being in the same neighborhood. By adding the Neighborhood variable to the model, perhaps the violation, if present, is somewhat mitigated.

Influential Points Analysis The Cook's D and outlier-leverage plots identify a few influential points (such as observations 1061 and 1190), but they they appear to be minor violations of the assumptions and were not removed (see Figures 12 and 13 in the appendix).

Model 2: Multiple Linear Regression (GrLivArea + FullBath)

Our second model is a multiple linear regression of SalePrice on GrLivArea and FullBath. This model seeks to predict home prices based on the living area and number of full bathrooms.

General Model

$$\log(\text{SalePrice}) = \beta_0 + \beta_1 \times \text{GrLivArea} + \beta_2 \times \text{FullBath} + \epsilon$$

Results

As with the simple linear model, we implemented Leave One Out Cross-Validation on our first multiple linear model to evaluate its performance. The results of our model are shown in Figure 14 in the appendix (as well as the full code on Github).

This results indicate that, on average, for every additional square foot of living area, the log of the sale price increases by 0.000458, and for each additional full bathroom, the log of the sale price increases by 0.1631, holding other variables constant. The intercept (11.08) represents the baseline log-sale price when both predictors are zero. This is obviously not practically feasible (as a house must have at least some living area, and it is unusual for a single family home to have no full bathrooms, though I have made do with some very small bathrooms!), thus it exists solely as a parameter in the model.

The adjusted R2 of this model is 0.5501, which means that 55% of the variance in log-transformed sale price can be explained by the size of the living area and the number of full bathrooms. This is a significant improvement over the simple linear model.

The linear regression of our equation is as follows:

$$\log(\text{SalePrice}) = 11.08 + 0.000458 \times \text{GrLivArea} + 0.1631 \times \text{FullBath}$$

Assumptions and Influential Points Analysis

Linearity The plot of residuals against fitted values shows no apparent pattern, indicating that the assumption of linearity is reasonably satisfied (see Figure 15 in the appendix).

Normality The Q-Q plot suggests that residuals follow a normal distribution, however some moderate deviation is observed at the tails (See Figure 16 in the appendix).

Variance The spread of residuals in the residuals plot (shown in Figure 17 in the appendix) is relatively uniform across fitted values, indicating that the assumption of constant variance (homoscedasticity) is met.

Independence No patterns or unusual are visible in the residuals plot (shown in Figure 17 in the appendix), which supports the assumption of independence.

Influential Points Analysis The Cook's D plot and the outlier-leverage plot identify a few points with high leverage or influence, such as observation 54. While these points do exceed the threshold for influence, we didn't think that they severely distorted the model, and we did not remove them (see Figures 17 and 18 in the appendix).

Model 3:

This model was found using the MASS package for step-wise feature selection targeting AIC. We will break from the above pattern and not include the formula for brevity's sake. In total there were 50 predictors in this model:

MSZoning, LotFrontage, LotArea, Street, Utilities, LotConfig, LandSlope, Neighborhood, Condition1, Condition2, BldgType, OverallQual, OverallCond, YearBuilt, YearRemodAdd, RoofStyle, RoofMatl, Exterior1st, ExterCond, Foundation, BsmtExposure, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, Heating, HeatingQC, CentralAir, Electrical, FirstFlrSF, SecondFlrSF, BsmtFullBath, FullBath, HalfBath, KitchenAbvGr, KitchenQual, TotRmsAbvGrd, Functional, Fireplaces, GarageType, GarageCars, GarageArea, GarageQual, GarageCond, WoodDeckSF, EnclosedPorch, ThirdSsnPorch, ScreenPorch, PoolArea, SaleType, SaleCondition

Assumptions and Influential Points Analysis

Figures 19-21 in the appendix

Linearity The plot of the residuals has a decently normal cloud around 0, but there is slight evidence of a bigger variance at the beginning and end of the plots.

Normality The Q-Q plot suggests that residuals follow a normal distribution moderately well, except for the tails of the distribution. This will be common theme going forward.

Variance The spread of the residuals as mention above provides some evidence of heteroscedasticity.

Independence We will assume that observations are independent enough to meet this assumption, however there is a possible cluster effect due to houses being in the same neighborhood. By adding the Neighborhood variable to the model, perhaps the violation, if present, is somewhat mitigated.

Influential Points Analysis There are numerous influential points in this model, however, this model predicts better than simpler models with less. "All models are wrong, but some are useful."

Appendix: Figures

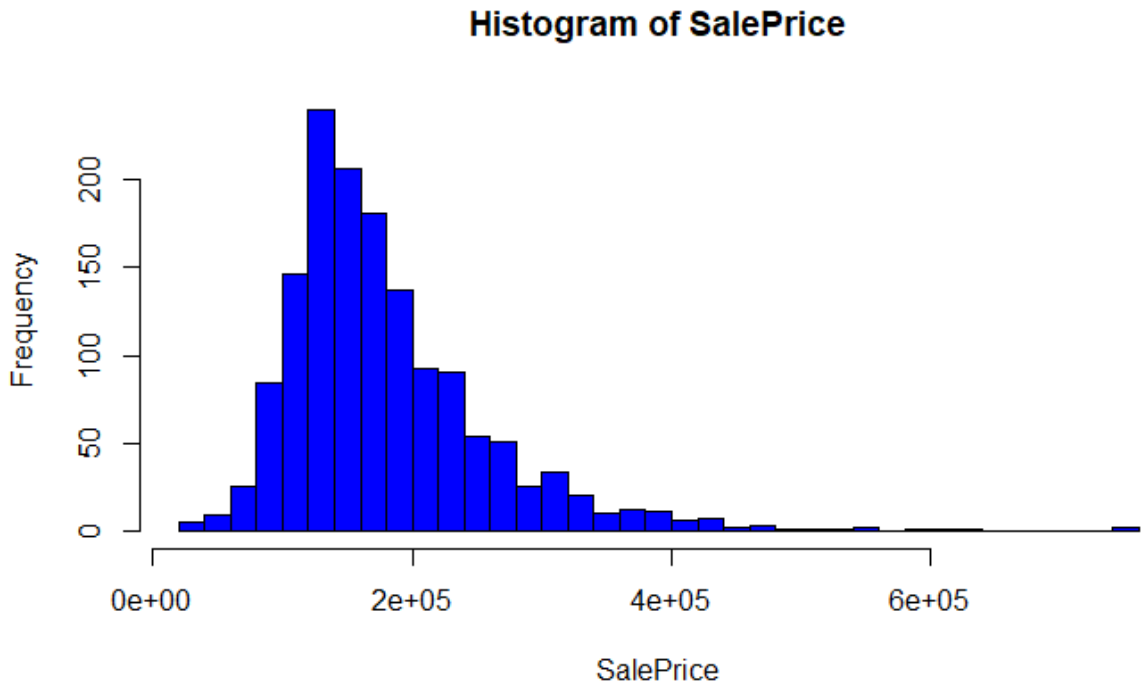

Figure 1: Figure 1: Scatterplot of SalePrice and GrLivAr

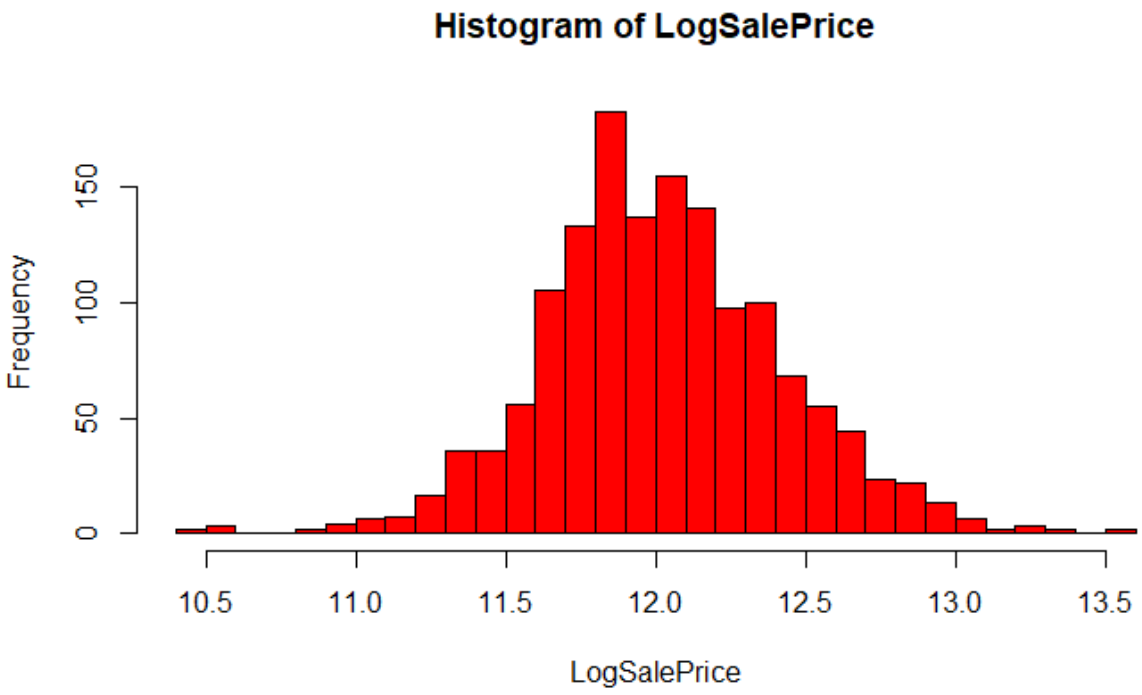

Figure 2: Figure 2: Scatterplot of SalePriceLog and GrLivAr

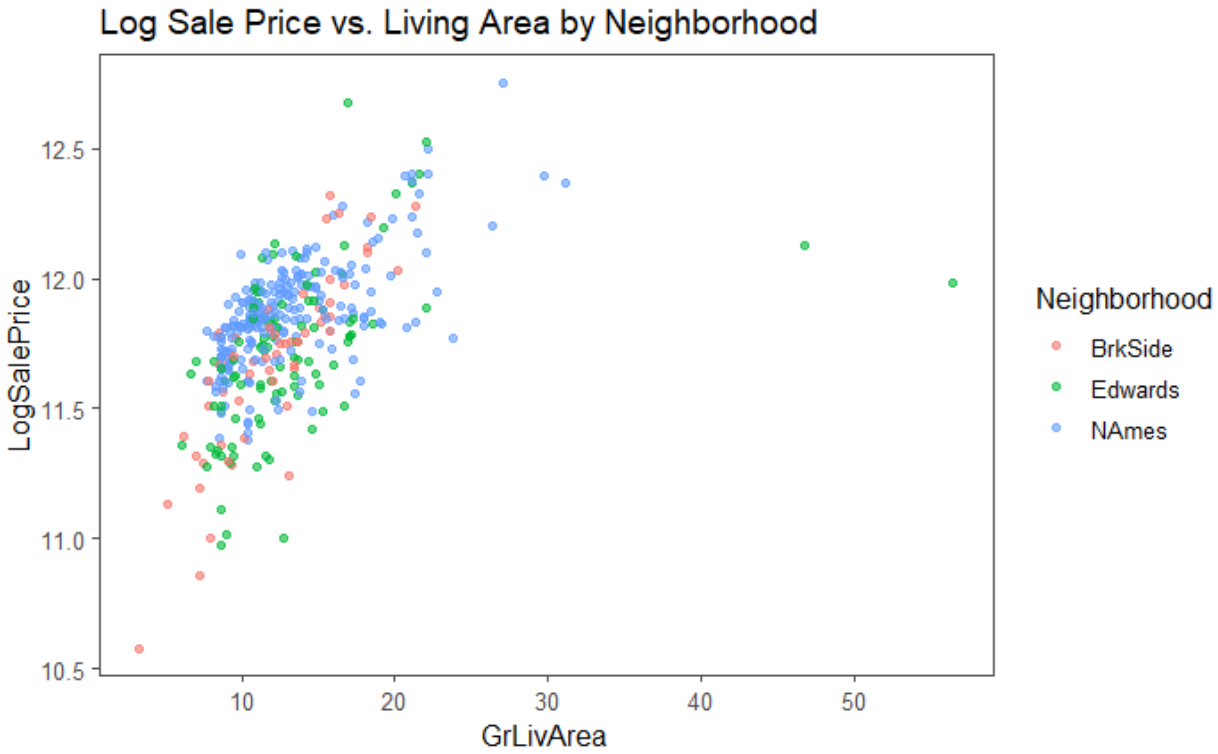

Figure 3: Figure 3: Scatterplot of SalePriceLog and GrLivAr by Neighborhood

Figure 4: Figure 4: Scatterplot of SalePriceLog and GrLivAr by Neighborhood

```
Call:
lm(formula = SalePriceLog ~ Neighborhood + GrLivArea:Neighborhood,
    data = data_filtered)

Residuals:
     Min       1Q   Median       3Q      Max
-0.71071 -0.11467  0.02085  0.11453  0.73596

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.079e+01  8.190e-02 131.773  < 2e-16 ***
NeighborhoodEdwards          2.339e-01  1.083e-01   2.160   0.0314 *
NeighborhoodNAmes            6.517e-01  9.205e-02   7.081 7.13e-12 ***
NeighborhoodBrkSide:GrLivArea 7.382e-04  6.486e-05  11.382  < 2e-16 ***
NeighborhoodEdwards:GrLivArea 5.387e-04  5.405e-05   9.966  < 2e-16 ***
NeighborhoodNAmes:GrLivArea   3.241e-04  3.059e-05  10.595  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1893 on 375 degrees of freedom
Multiple R-squared:  0.5271,    Adjusted R-squared:  0.5208
F-statistic: 83.61 on 5 and 375 DF,  p-value: < 2.2e-16
```

Figure 5: Figure 5: Summary of Descrtiptive Model

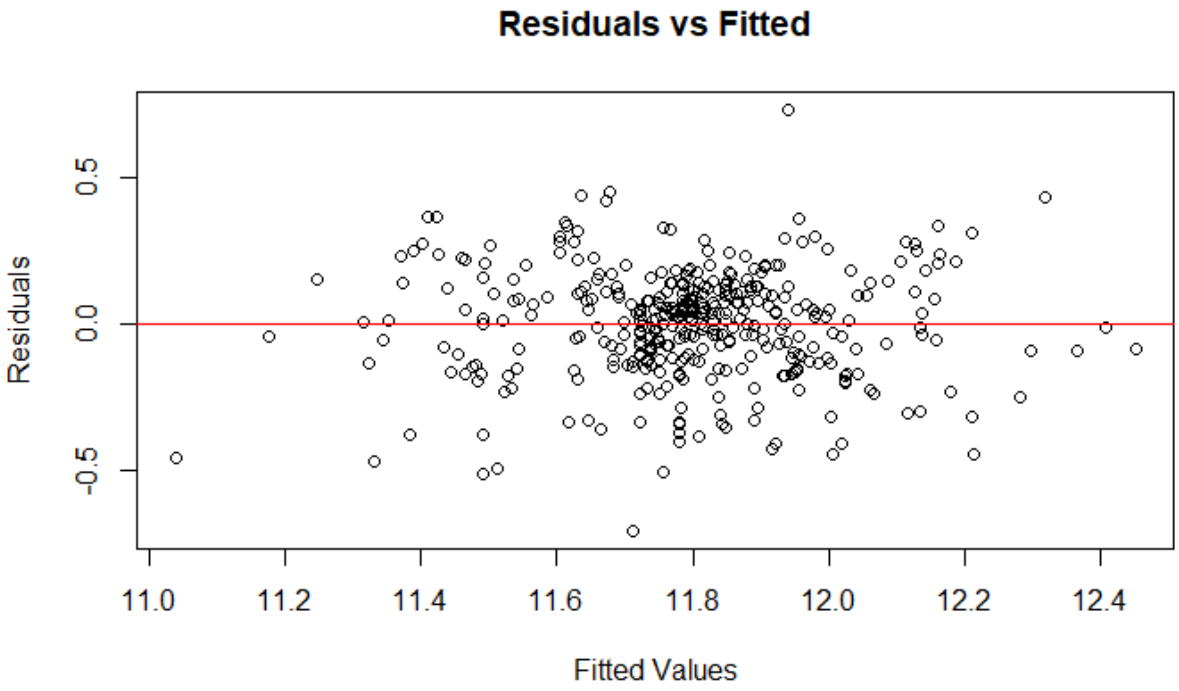

Figure 6: Figure 6: Residuals vs Fitted, Descriptive Model

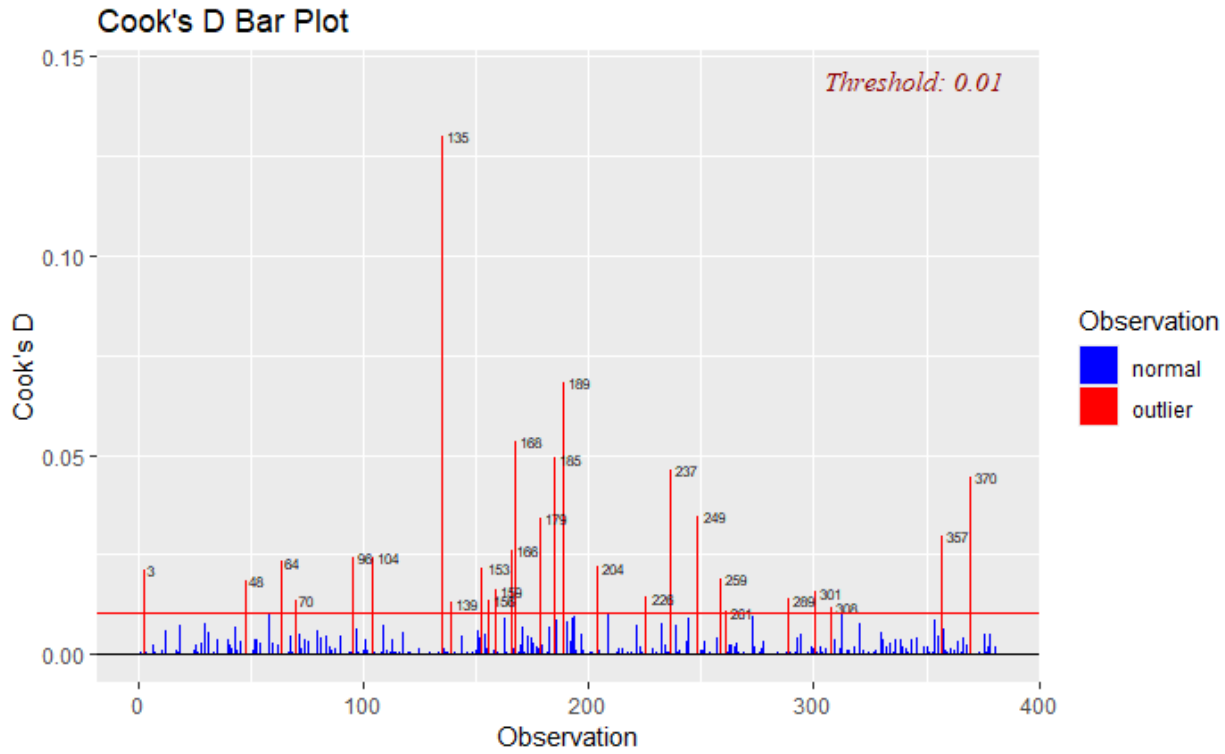

Figure 7: Figure 7: Cook's D Plot for Descriptive Model

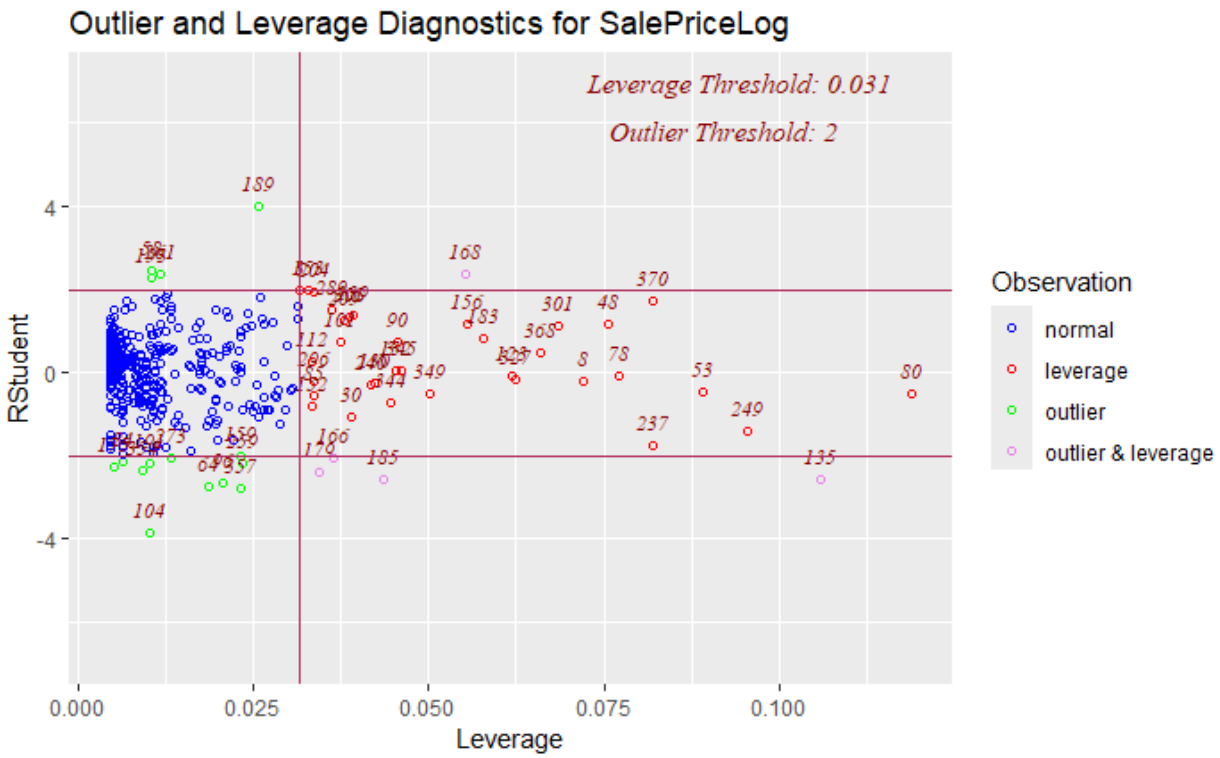

Figure 8: Figure 8: Outlier-Leverage Plot for Descriptive Model

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.68065 -0.15803  0.01765  0.17932  1.07582

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.144e+01  1.928e-02  593.50   <2e-16 ***
GarageArea  1.236e-03  3.725e-05   33.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3017 on 1456 degrees of freedom
Multiple R-squared:  0.4305,    Adjusted R-squared:  0.4301
F-statistic:  1101 on 1 and 1456 DF,  p-value: < 2.2e-16

Linear Regression

1458 samples
   1 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 1457, 1457, 1457, 1457, 1457, 1457, ...
Resampling results:

  RMSE      Rsquared   MAE
  0.302109  0.4283626  0.2252561

Tuning parameter 'intercept' was held constant at a value of TRUE
AIC: 647.785041926771
BIC: 663.639504664477
PRESS: 133.071433024278
```

Figure 9: Figure 9: Summary of Simple Linear Model

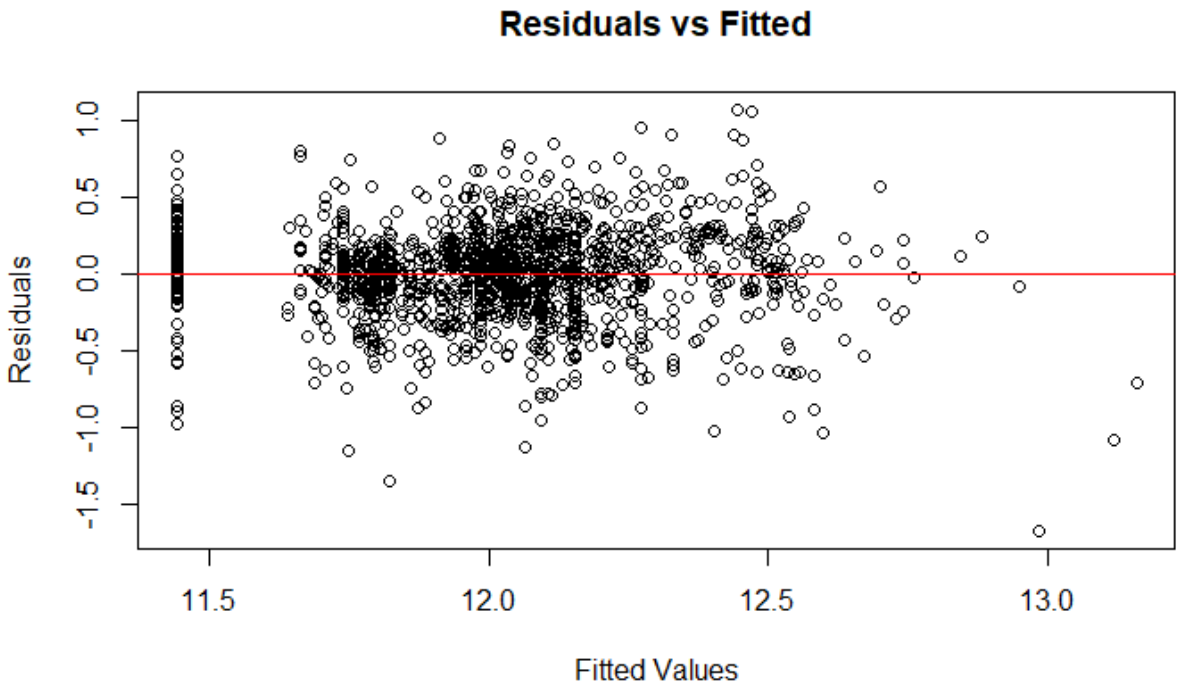

Figure 10: Figure 10: Residuals vs Fitted for Simple Linear Model

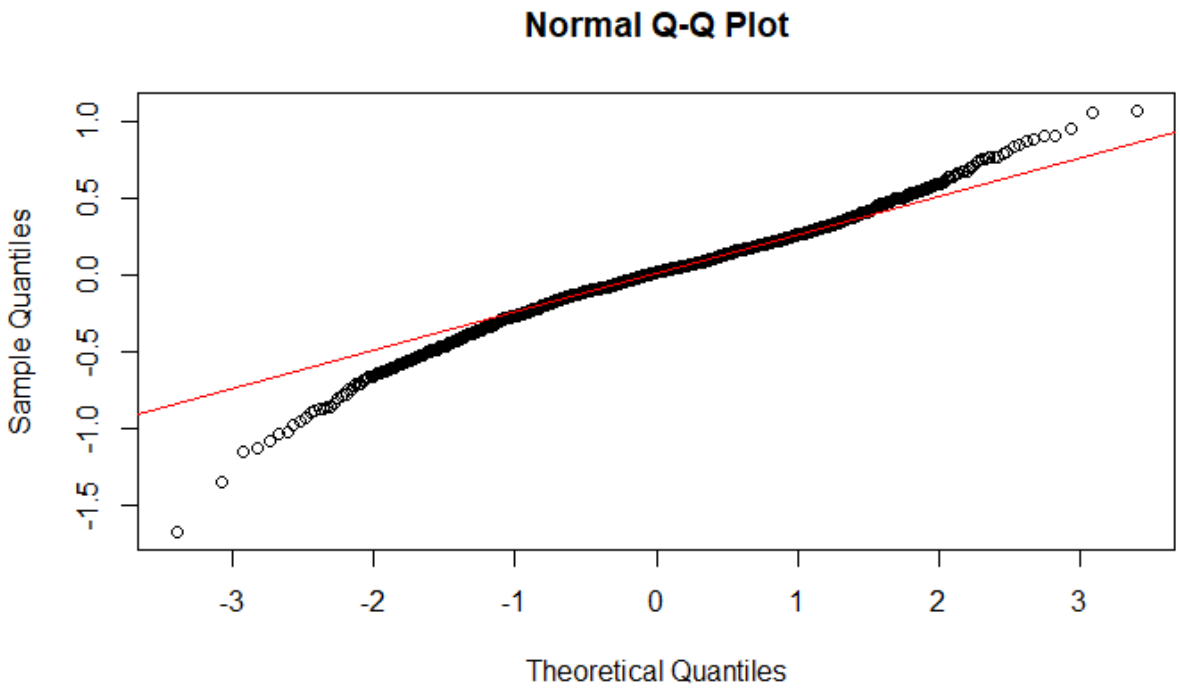

Figure 11: Figure 11: Q-Q Plot for Simple Linear Model

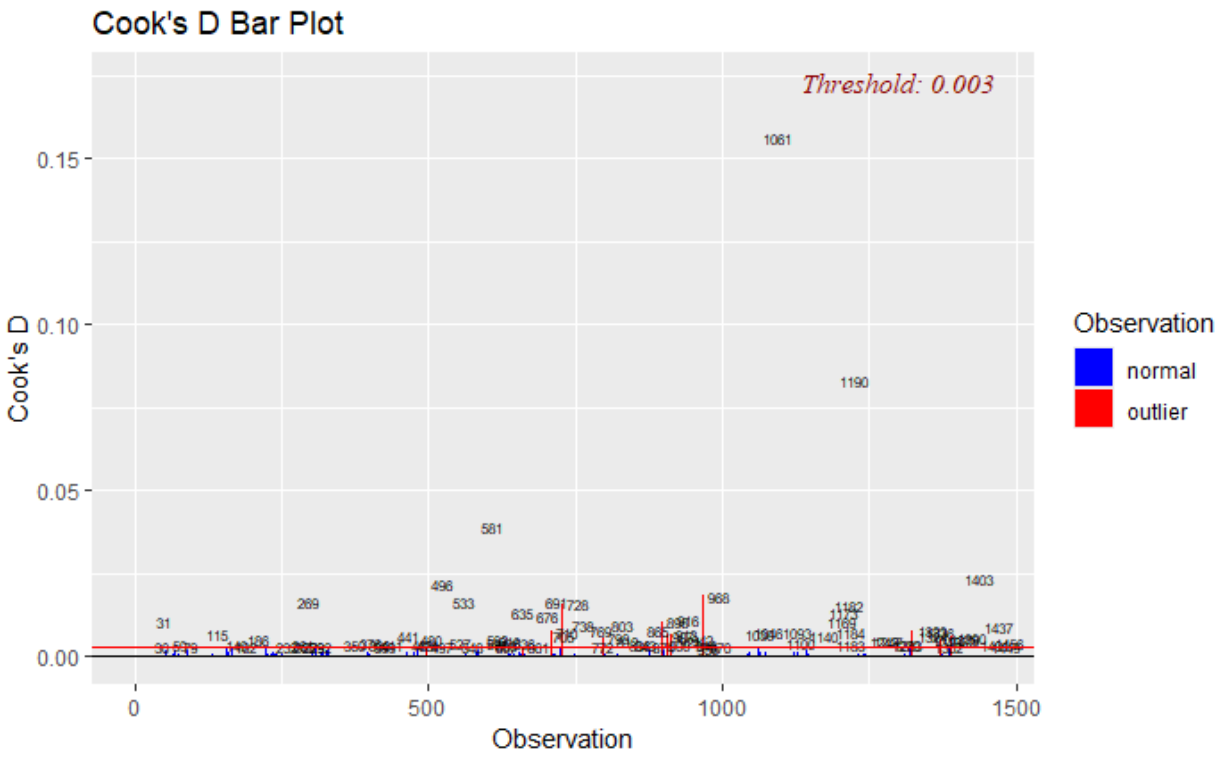

Figure 12: Figure 12: Cook's D Plot for Simple Linear Model

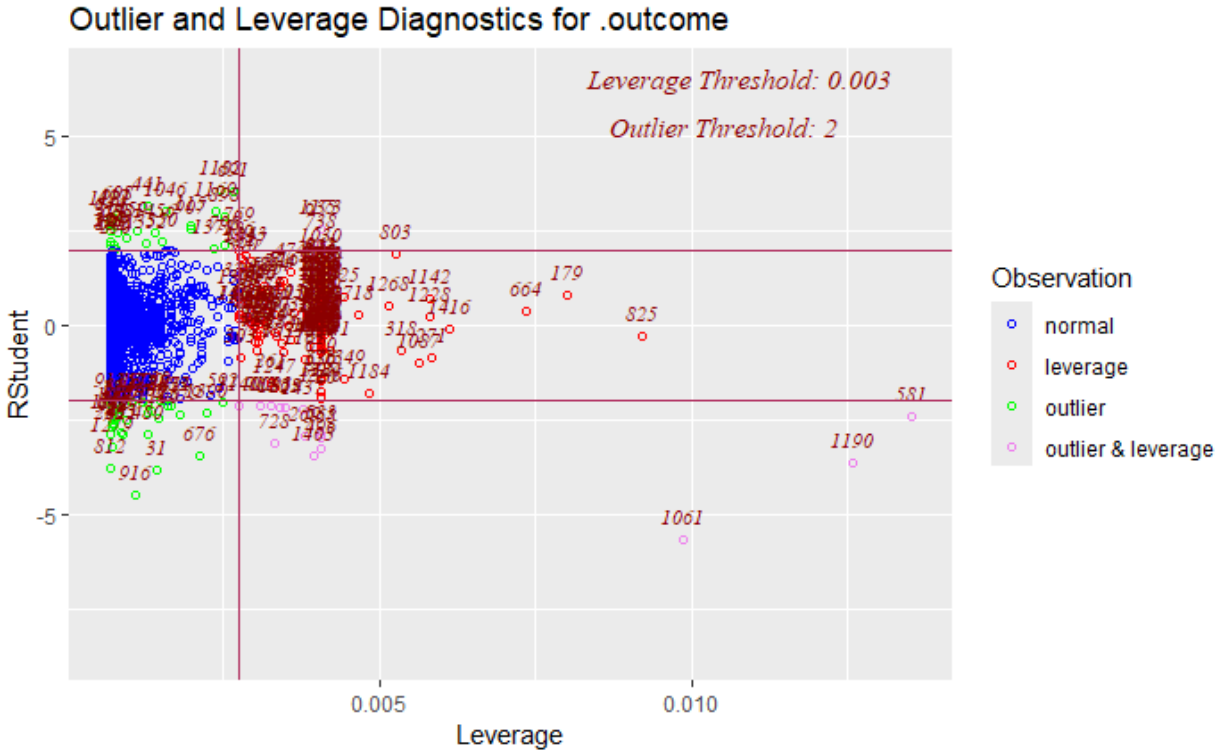

Figure 13: Figure 13: Outliers-Leverage Plot for Simple Linear Model

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24681 -0.12426  0.02505  0.15296  0.94020

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.108e+01  2.353e-02 470.711   <2e-16 ***
GrLivArea   4.580e-04  1.787e-05  25.636   <2e-16 ***
FullBath    1.631e-01  1.650e-02   9.884   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2666 on 1455 degrees of freedom
Multiple R-squared:  0.5558,    Adjusted R-squared:  0.5551
F-statistic: 910.1 on 2 and 1455 DF,  p-value: < 2.2e-16

Linear Regression

1458 samples
   2 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 1457, 1457, 1457, 1457, 1457, 1457, ...
Resampling results:

  RMSE      Rsquared  MAE
  0.266976  0.553586  0.1968536

Tuning parameter 'intercept' was held constant at a value of TRUE
AIC: 287.63780388207
BIC: 308.777087532345
PRESS: 103.920677063249
```

Figure 14: Figure 14: Summary of GrLivArea + FullBath Model

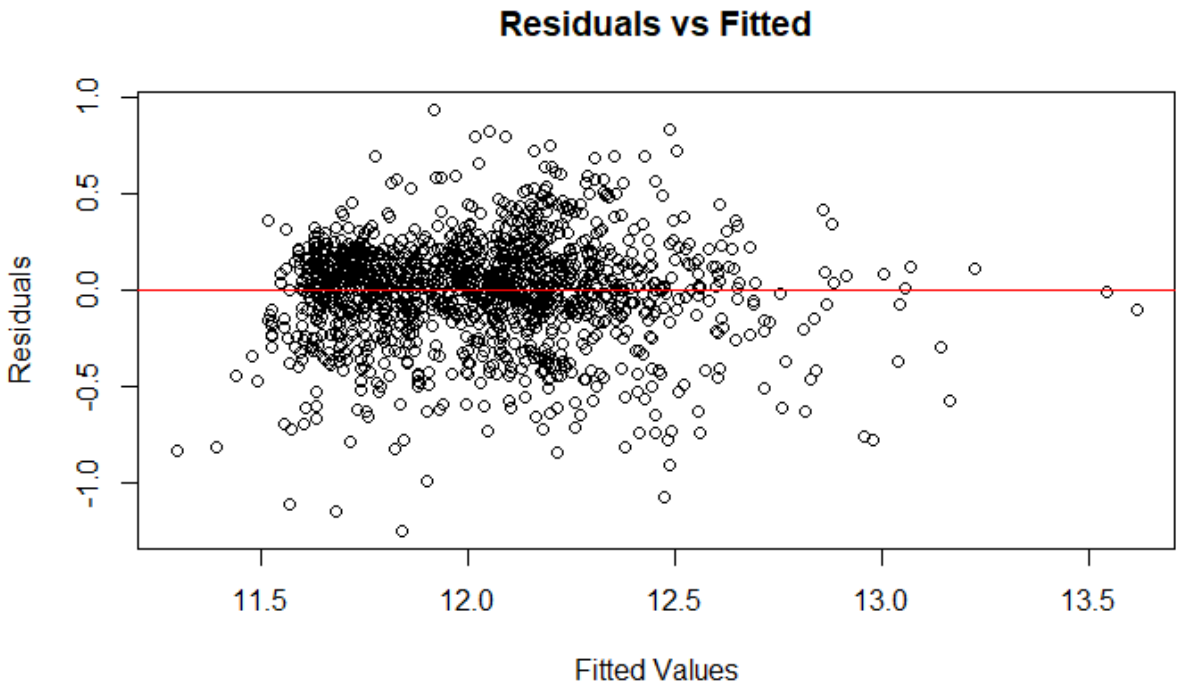

Figure 15: Figure 15: Residuals vs Fitted for GrLivArea + FullBath Model

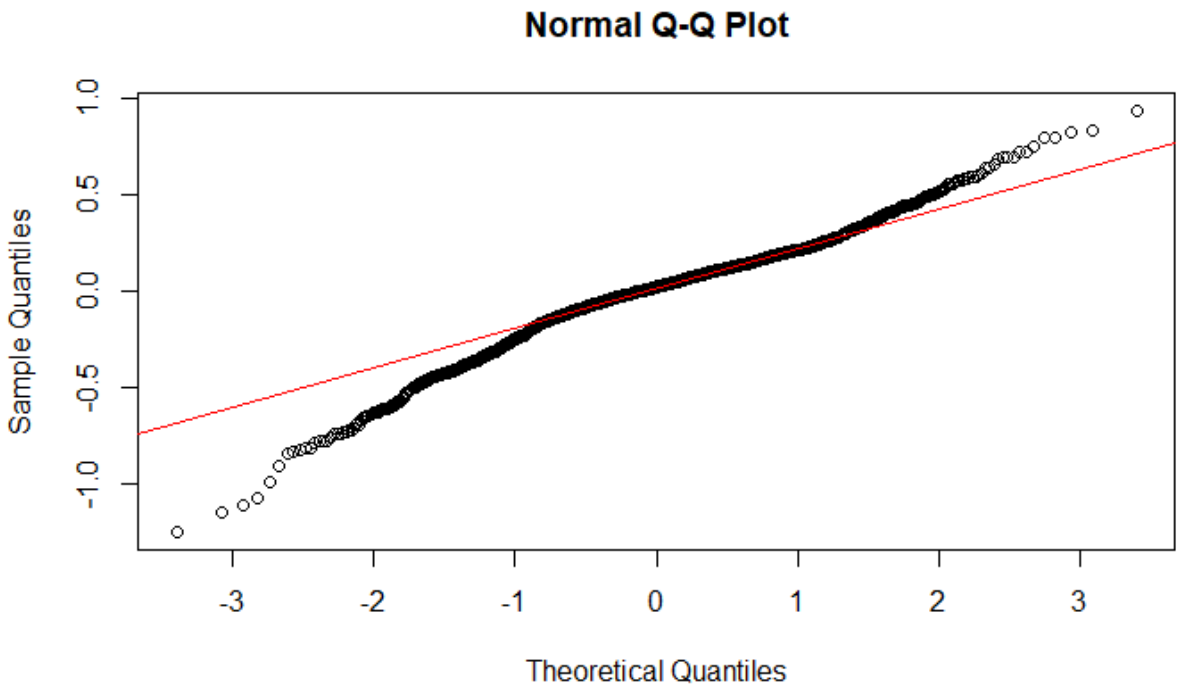

Figure 16: Figure 16: Q-Q Plot for GrLivArea + FullBath Model

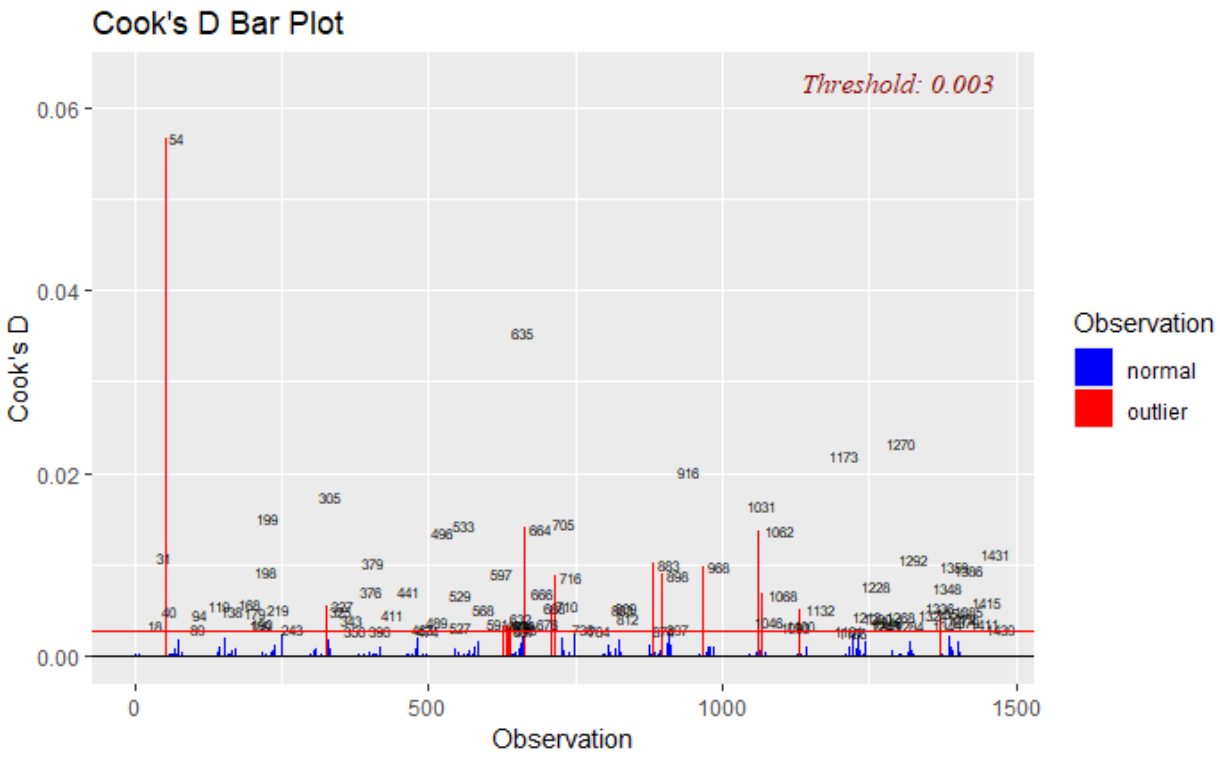

Figure 17: Figure 17: Cook's D Plot for GrLivArea + FullBath Model

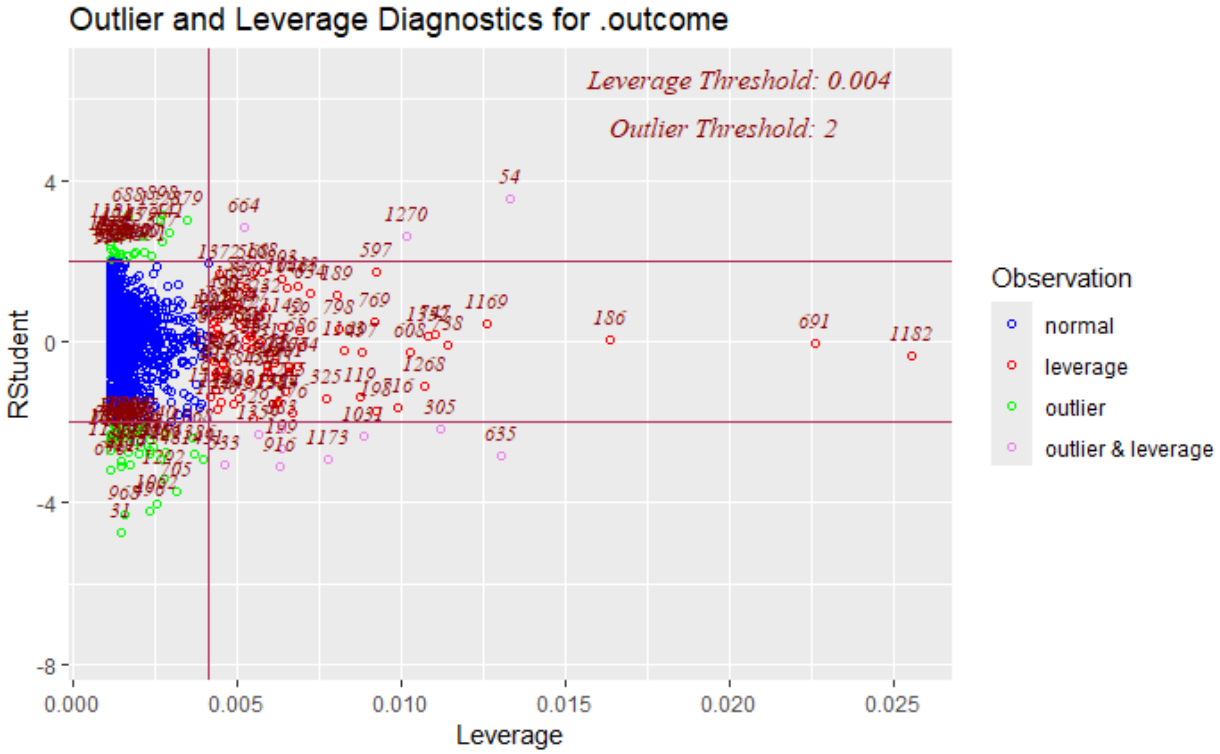

Figure 18: Figure 18: Outliers-Leverage Plot for GrLivArea + FullBath Model

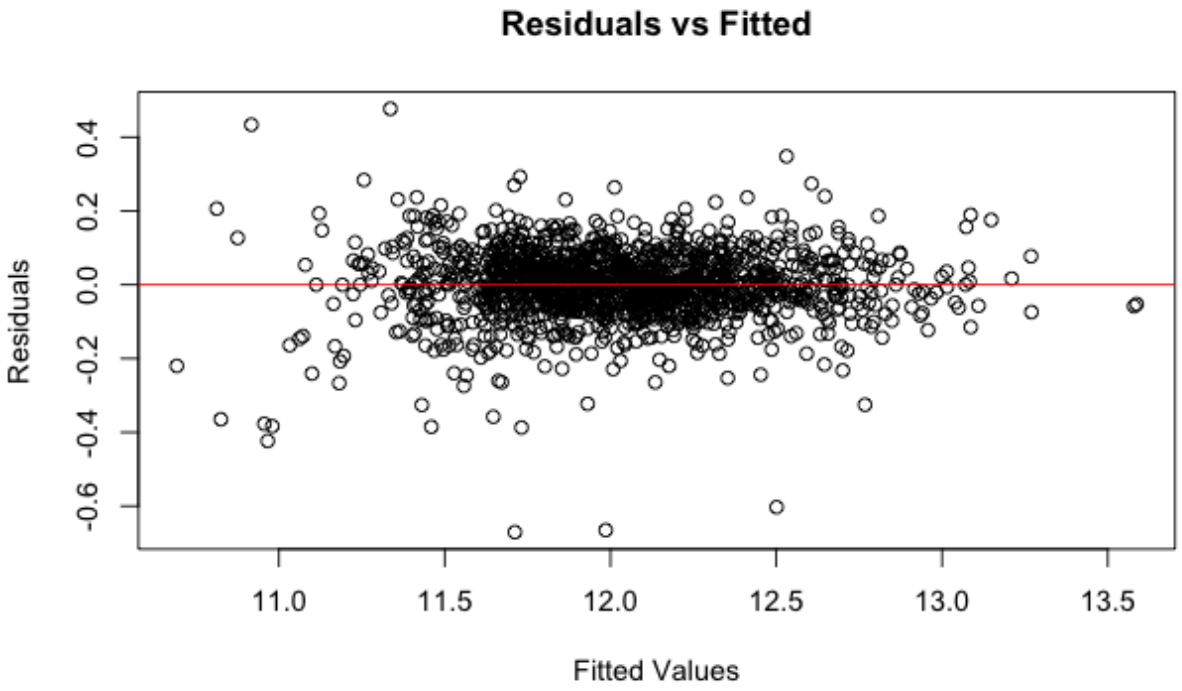

Figure 19: Figure 19: Model 3 Scatter Plot of Residuals

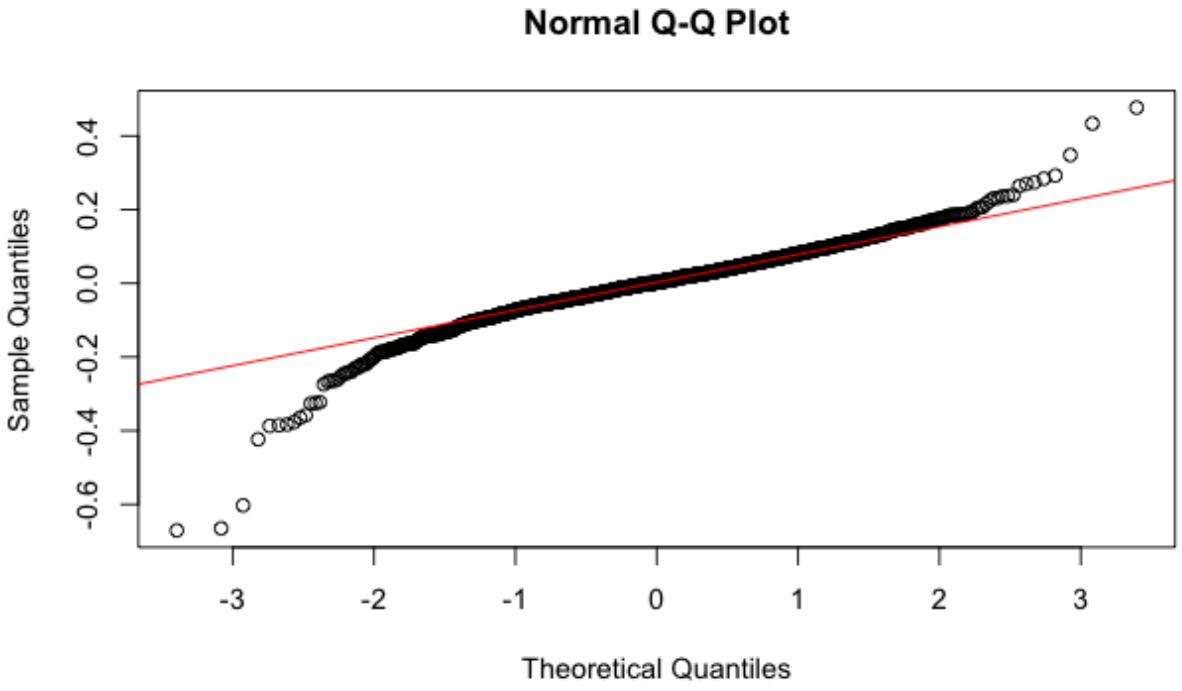

Figure 20: Figure 20: Model 3 QQ Plot of Residuals

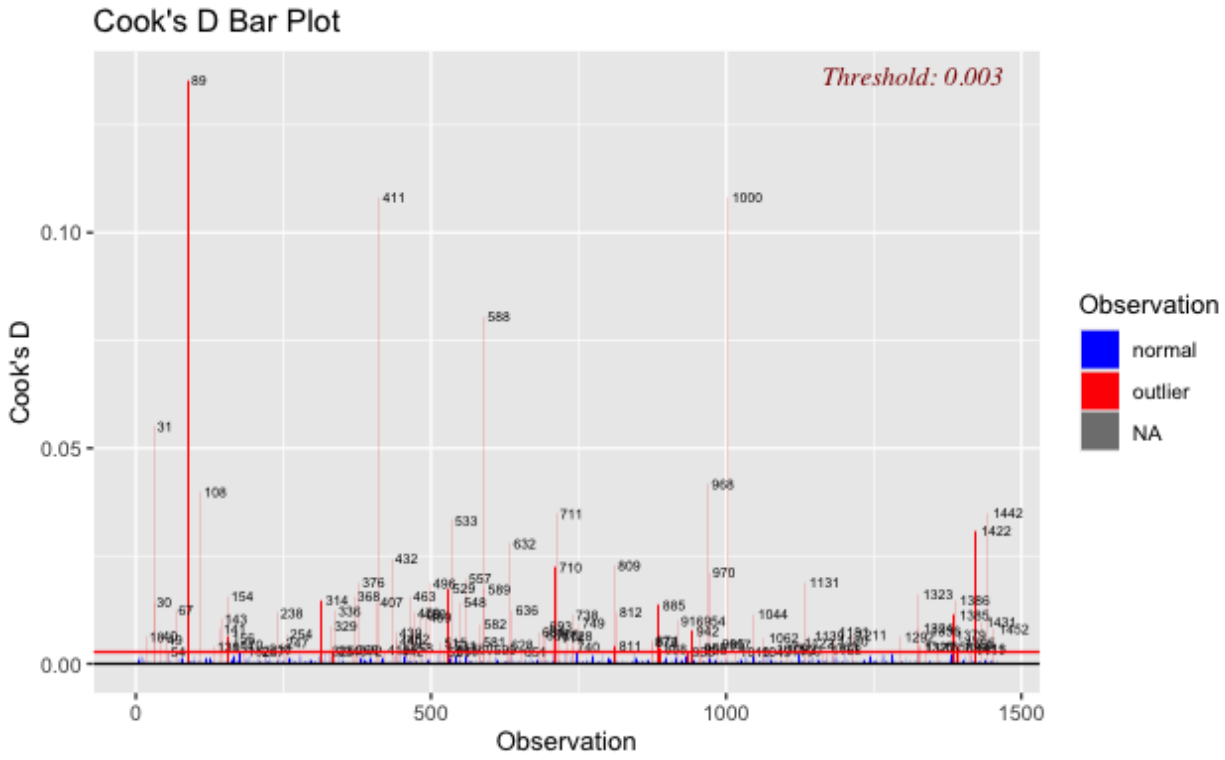

Figure 21: Figure 21: Model 3 Cook's D