Effect on House Prices

Victoria Blante, Ruben Cabrera, and Thao Tran

May 9, 2023

May 10, 2023

Abstract

For this project, we worked with the Ames Housing data set, containing information on houses in Ames, Iowa. The goal was to find significant predictors and build a multiple linear regression model to predict the sale price of homes in Ames. The data set originally had 82 variables, but through model selection and further analysis, our final modal ended with four predictors; Year_Built, Gr_Liv_Area, TotRoms_AbvGrd, and Lot_Area. Through our exploration and analysis of the data, we decided to use a log transformation on the predictor Sale_Price along with one of the predictors Lot_Area. We found that these transformations helped the data better to meet the assumptions of a multiple linear regression model.

Problem and Motivation

The AmesHousing dataset was created to address a problem faced by real estate agents and appraisers in Ames, Iowa. In the past, they used a set of standardized values to estimate home prices, which did not accurately reflect the local housing market. As a result, they needed a more precise and comprehensive dataset that included a wider range of variables that influence house prices in Ames.

The motivation behind creating the AmesHousing dataset was to provide a more accurate and up-to-date tool for real estate agents and appraisers to estimate the value of homes in Ames. By including a comprehensive set of variables such as the age of the house, the number of bedrooms and bathrooms, the size of the lot and living area, and other features, the dataset enables better predictions of home values based on local market trends.

Furthermore, the dataset has also been widely used in machine learning and statistical modeling research to investigate the relationship between various factors and housing prices, providing insights into the real estate market in Ames and beyond.

Data Description

For our analysis, we used the Ames Housing data set, which contains information about 2,930 Ames, Iowa properties sold between 2006 and 2010. This data set was found in an R package with the same title. The columns include information related to the sale price, house characteristics, location, and lot information. We used sale price as our response variable and worked through some model selections to choose our predictors. This data set has 82 columns with 23 nominal, 14 discrete, 20 continuous variables, and two observational identifiers.

Question of interest

We had some questions of interest before doing any analysis after we explored the dataset. Here are some questions that came up:

1. What are the 4 most important factors that affect the Price in that area?
2. What would be the best opportunities to buy a house under the market price in Ames?
3. What would be the worst properties within that database to buy?
4. How the age of the house affects the price?
5. What year in the dataset had more good deal houses?

Regression Analysis, Results, and Interpretation

We created a subset of 9 variables from the dataset so it is easier to explain and perform the regression analysis. We first made a histogram and added-variable plots to see which variables are statistically significant, have a statistical relationship with the response, and work for a multiple linear regression model.

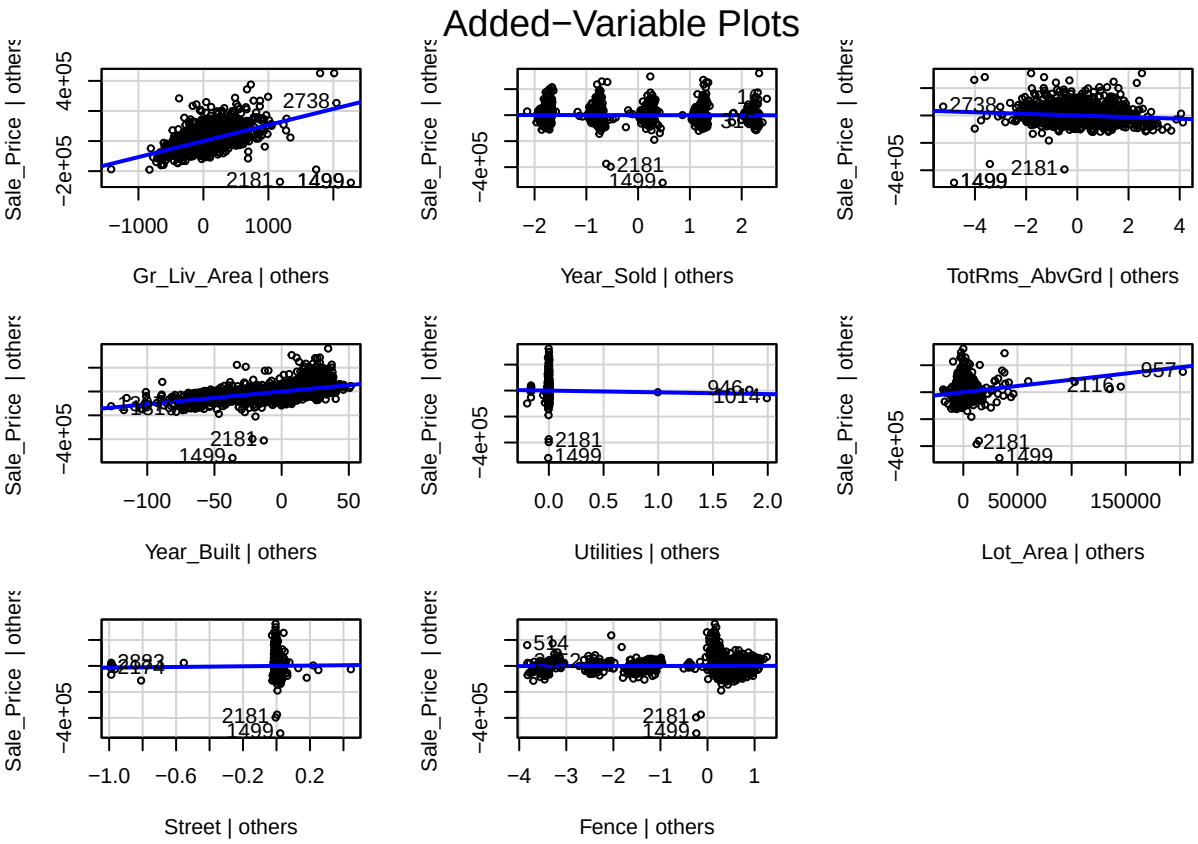

Figure 1: The partial regression plots of 8 predictors from the subset Ames Housing dataset in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

The partial regression plots show the relationship between a response variable, Sale_Price, and eight other predictors. These plots detect non-linear relationships (horizontal lines residuals around zero) and outliers in the data, Year_Sold, Utilities, Street, and Fence. Other variables, Gr_Liv_Area, TotRms_AbvGrd, Year_Built, and Lot_Area, have significant relationships with the response. We then computed the full model to confirm what we found from those plots.

```
##
## Call:
## lm(formula = Sale_Price ~ ., data = ames2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -519121  -25798   -2793   18721  323686
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.060e+06  1.296e+06  -0.818    0.413
## Gr_Liv_Area    1.068e+02  3.007e+00  35.515  < 2e-16 ***
## Year_Sold     -5.025e+02  6.442e+02  -0.780    0.435
## TotRms_AbvGrd -5.925e+03  9.241e+02  -6.412 1.67e-10 ***
```

```
## Year_Built     1.067e+03  2.973e+01  35.883  < 2e-16 ***
## Utilities     -1.345e+04  1.559e+04  -0.863    0.388
## Lot_Area       9.229e-01  1.138e-01   8.108 7.48e-16 ***
## Street         1.339e+04  1.370e+04   0.977    0.328
## Fence          3.433e+02  7.954e+02   0.432    0.666
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45840 on 2921 degrees of freedom
## Multiple R-squared:  0.6716, Adjusted R-squared:  0.6707
## F-statistic: 746.7 on 8 and 2921 DF,  p-value: < 2.2e-16
```

Table 1: The summary table of the full model with eight predictors and one response in Ames, Iowa, between 2006 and 2010 (data obtained from the R package).

The summary of the full model also confirms that Year_Sold, Utilities, Street, and Fence are not statistically significant because their p-values 0.435, 0.388, 0.328, and 0.666, respectively, and are all larger than the standard values 0.001, 0.01, and 0.05 of the null hypothesis. Next, we performed the variable selection to choose which variables could be used for the final model using Forward Stepwise Selection with the AIC method.

```
##
## Call:
## lm(formula = Sale_Price ~ Gr_Liv_Area + Year_Built + Lot_Area +
##     TotRms_AbvGrd, data = na.omit(ames2))
##
## Coefficients:
##   (Intercept)    Gr_Liv_Area     Year_Built       Lot_Area  TotRms_AbvGrd
##    -2.063e+06      1.070e+02      1.071e+03      9.010e-01     -5.902e+03
```

Table 2: Forward Stepwise Selection with AIC table computed with nine predictors in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

The results indicate that the model assumptions have been met and predictor variables are not highly correlated with each other, which is a good sign for not having multicollinearity and overfitting the data. Therefore, this can improve the regression model's stability, reliability, and interpretability. We used these five variables for our reduced model and refitted the model.

```
##
## Call:
## lm(formula = Sale_Price ~ Gr_Liv_Area + Year_Built + Lot_Area +
##     TotRms_AbvGrd, data = ames3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -518992  -25792   -2738   18837  322425
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.063e+06  5.745e+04 -35.913  < 2e-16 ***
## Gr_Liv_Area    1.070e+02  3.002e+00  35.650  < 2e-16 ***
## Year_Built     1.071e+03  2.921e+01  36.683  < 2e-16 ***
## Lot_Area       9.010e-01  1.123e-01   8.020 1.51e-15 ***
## TotRms_AbvGrd -5.902e+03  9.237e+02  -6.389 1.93e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 45830 on 2925 degrees of freedom
## Multiple R-squared:  0.6713, Adjusted R-squared:  0.6708
## F-statistic:  1493 on 4 and 2925 DF,  p-value: < 2.2e-16
```

Table 3: The summary table of the reduced model with five variables in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

Table 3 represents all the significant variables after checking and selecting variables. We then moved to the next step of checking model assumptions for the reduced linear regression model using the diagonal plots below.

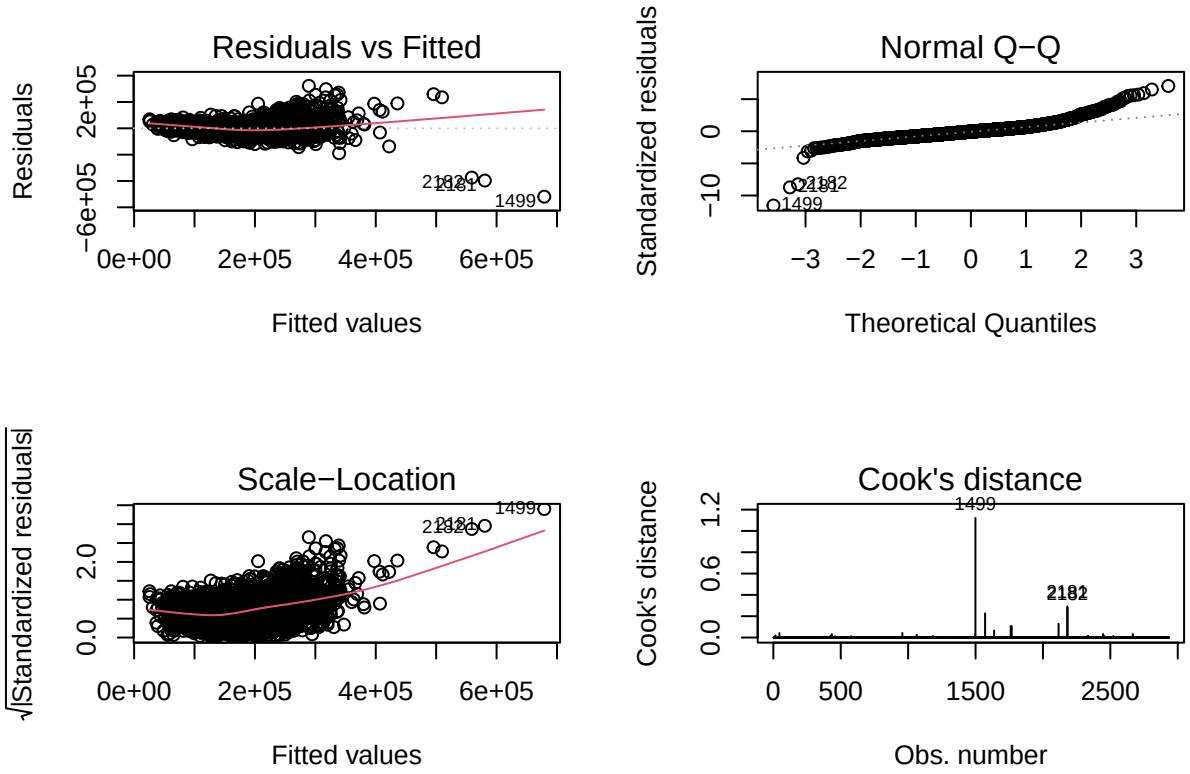

Figure 2: Diagnostic plots of the reduced model in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

There are some violations in linearity, normality, and outliers issues in these residual plots in fFigure2. There is non-linearity in the residual vs. fitted plot since the line is a bit curved. The homoscedasticity seems roughly constant across the range of fitted values though there are outliers.

The QQ plot shows the data is not normally distributed since most of the data points are scattered around the linear regression line, but there are outliers and a heavy tail on the right side.

The Scale-Location and Cook's distance plots also confirm those violations. Thus, the transformation is needed for this model. Before we transform the model, we would like to investigate the outliers to determine whether we should keep or remove them from the model.

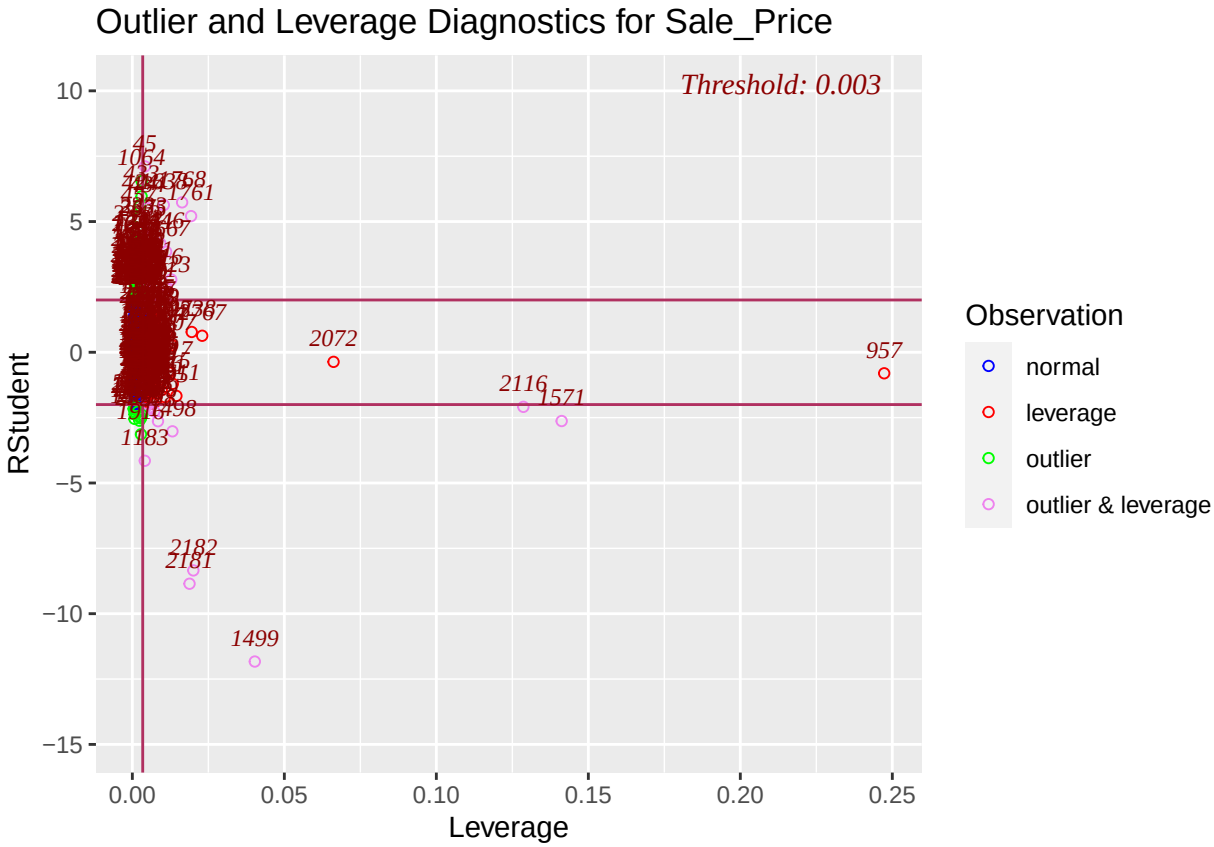

Figure 3: Outlier and Leverage Diagnostics for Sale_Price in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

Figure 3 shows some bad leverage data points far from the horizontal lines, so we investigated those outliers by looking at the dataset.

```
## # A tibble: 6 x 5
##   Sale_Price Gr_Liv_Area Year_Built Lot_Area TotRms_AbvGrd
##        <int>       <int>      <int>    <int>         <int>
## 1     160000        5642       2008    63887            12
## 2     183850        5095       2008    39290            15
## 3     184750        4676       2007    40094            11
## 4     148000        1072       2005     3675             5
## 5     171500        1408       2005     7023             6
## 6     140000        1072       2005     3675             5
```

Table 3: Data rows for outlier points obtained from the original dataset in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

The first three rows of Table 3 show where those outliers are located, and the last three rows represent those data points that are normally distributed in the dataset. There are unusual things about those outliers are that prices of those houses seem to reasonable based on the information provided in the dataset, like the Year_Built and Gr_Liv_Area; however, Lot_Area and TotRms_AbvGrd data appears to be odd because the square feet of lot area and the total room above the ground are way larger compared to other data points. Still, the prices are similar to other houses. This might be due to typing mistake, or those houses were not located nearby the rest of the homes in this research. However, we decided to keep those outliers because they represent actual observations that are important to the analysis and do not affect the sale price.

After dealing with the outliers, we used other approaches to fix or improve problems with non-linearity and normality from diagnostic checks.

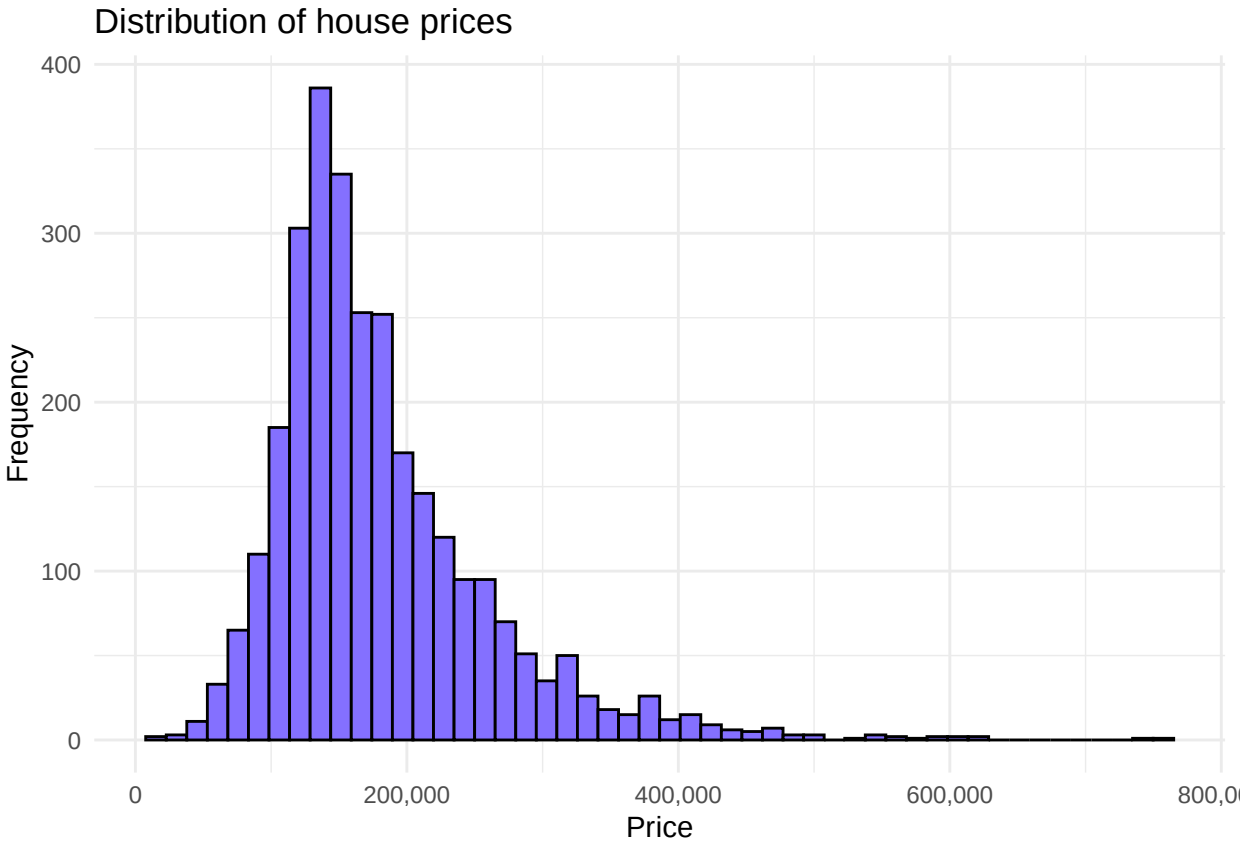

Figure 4: The histogram of the response, Sale_Price in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

The histogram of the response variable is skewed to the right, which indicates the data is not normally distributed, so we decided to use the log transformation for the response variable.

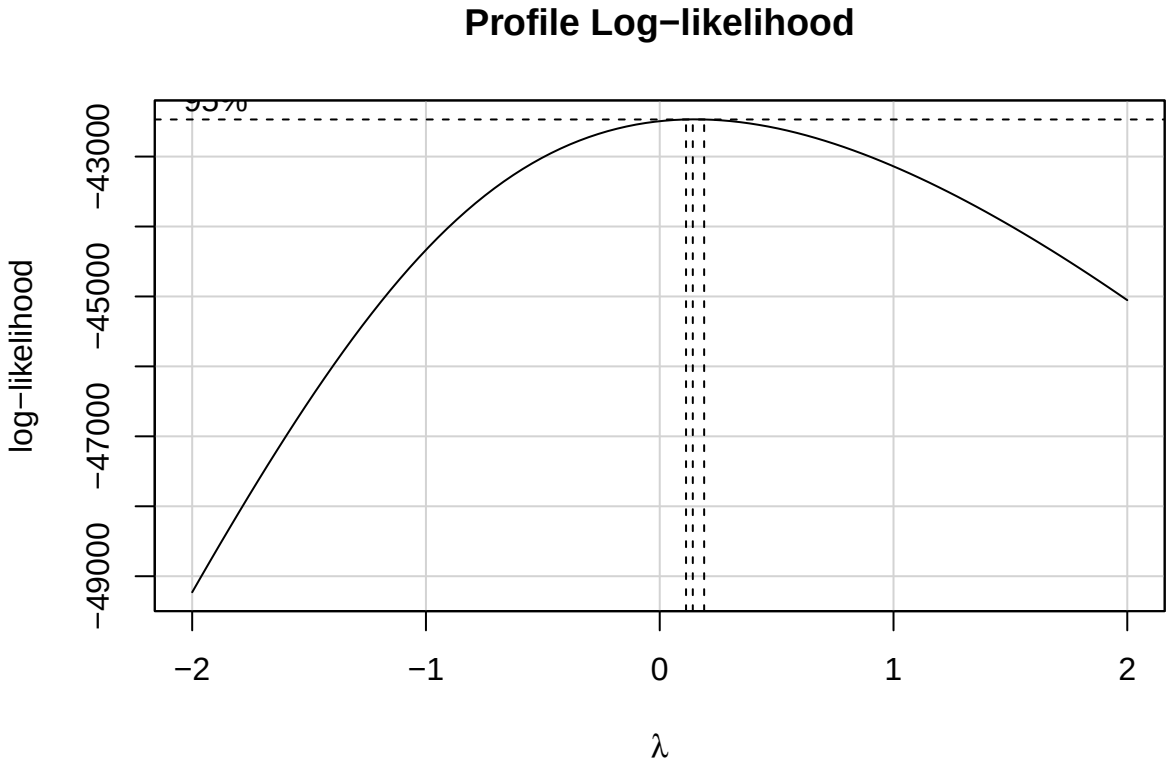

Figure 5: Profile Log-Likelihood plot of the response, Sale_Price in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

Figure 5 also confirms that the response transformation is necessary because the estimate (lambda) is near 0, so we rounded it to 0, and log transformation is an appropriate approach. After the transformation, we plotted another histogram to compare how the log transformation helped improve the model.

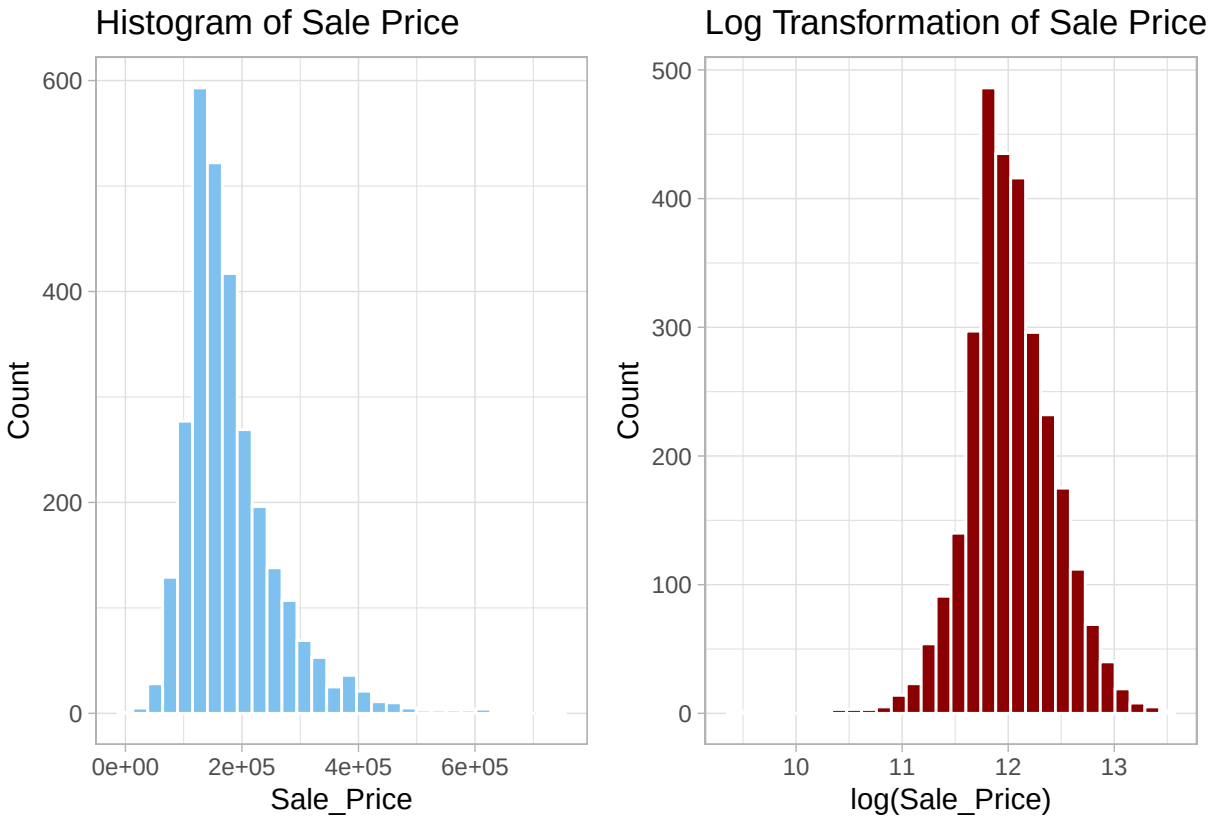

Figure 6: The side-by-side histogram of the response before and after the log transformation in Ames, Iowa, between 2006 and 2010 (data obtained from the R package).

Figure 6 shows that the bars now have a bell shape instead of skewing to the right like the histogram on the left side, indicating the data is more normally distributed in the histogram in the right hand. We also made another scatterplot matrix to view how the model improved after the response transformation.

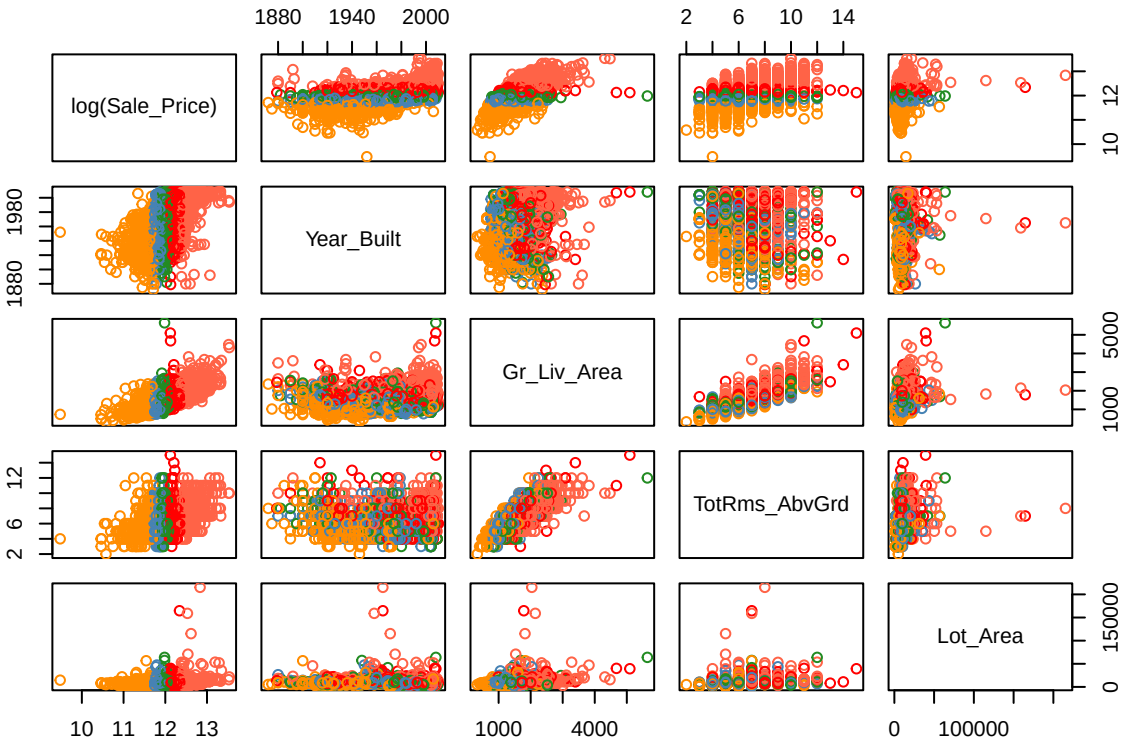

Figure 7: The scatterplot matrix of the full model with log(Sale_Price) and four predictors held fixed in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

Figure 7 shows that there seems to be multicollinearity in the predictor, TotRms_AbvGrd, and outliers in the Lot_Area variable. Therefore, we checked the correlation to ensure there would be no issue with multicollinearity in this model. We made the coefficient matrix plot below.

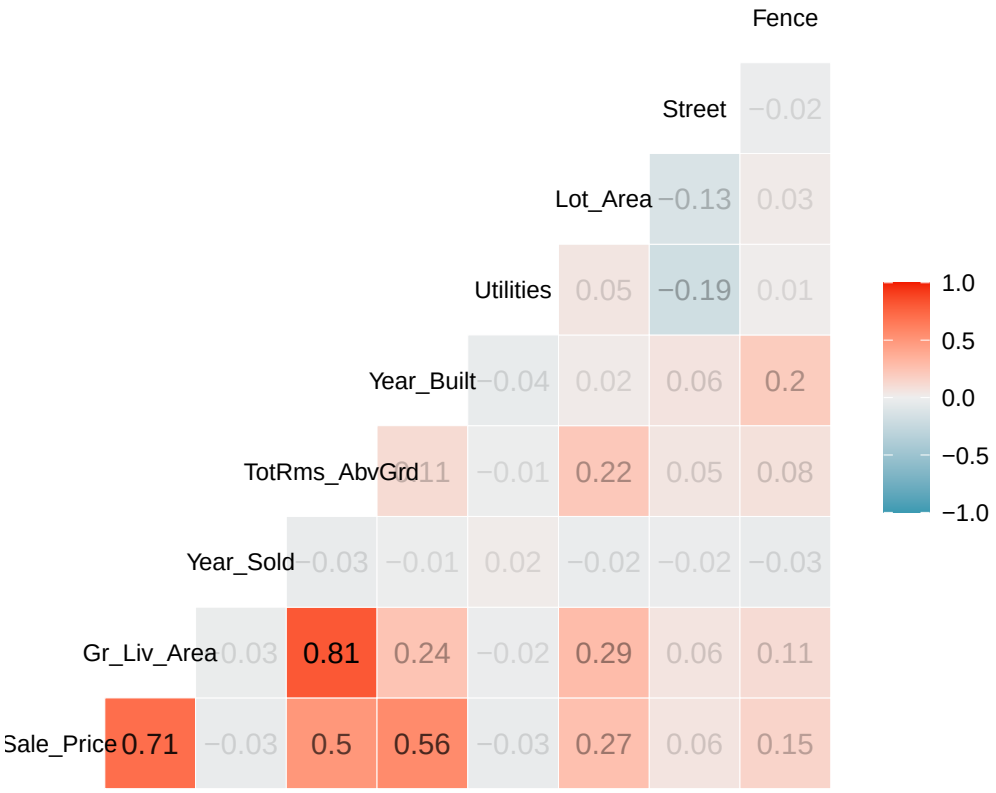

Figure 8: The coefficient matrix plot of all reduced models between the response and predictors in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

The darker shades of red in Figure 8 indicate a stronger positive correlation, and the darker shades of blue represent a stronger negative correlation. On the other hand, lighter shades show values closer to zero (weaker linear relation). All predictors have a strong relationship with the outcome variable and are significant predictors in the model.

```
##    Year_Built   Gr_Liv_Area TotRms_AbvGrd      Lot_Area
##          1.09          3.21          2.94          1.09
```

Table 4: Variance Inflation Factor (VIF) predictors in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

We double-check the multicollinearity using the VIF method, and Table 4 shows the test results. The values from the VIF test for all predictors are less than 5, indicating no multicollinearity problems.

```
##            df     AIC
## lm_logfit1  6 -517.36
## lm_logfit2  6 -676.36
```

Table 5: AIC table of two models with and without log(Lot_Area) in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

The AIC results from Table5 with log transformation for Lot_Area on the second row is -676.3639, smaller than the model without log(Lot_Area) is -515,3620, which indicates that the log transformation was needed for this model. We made another scatterplot matrix to look at these variables again after transforming Sale_Price and Lot_Area variables.

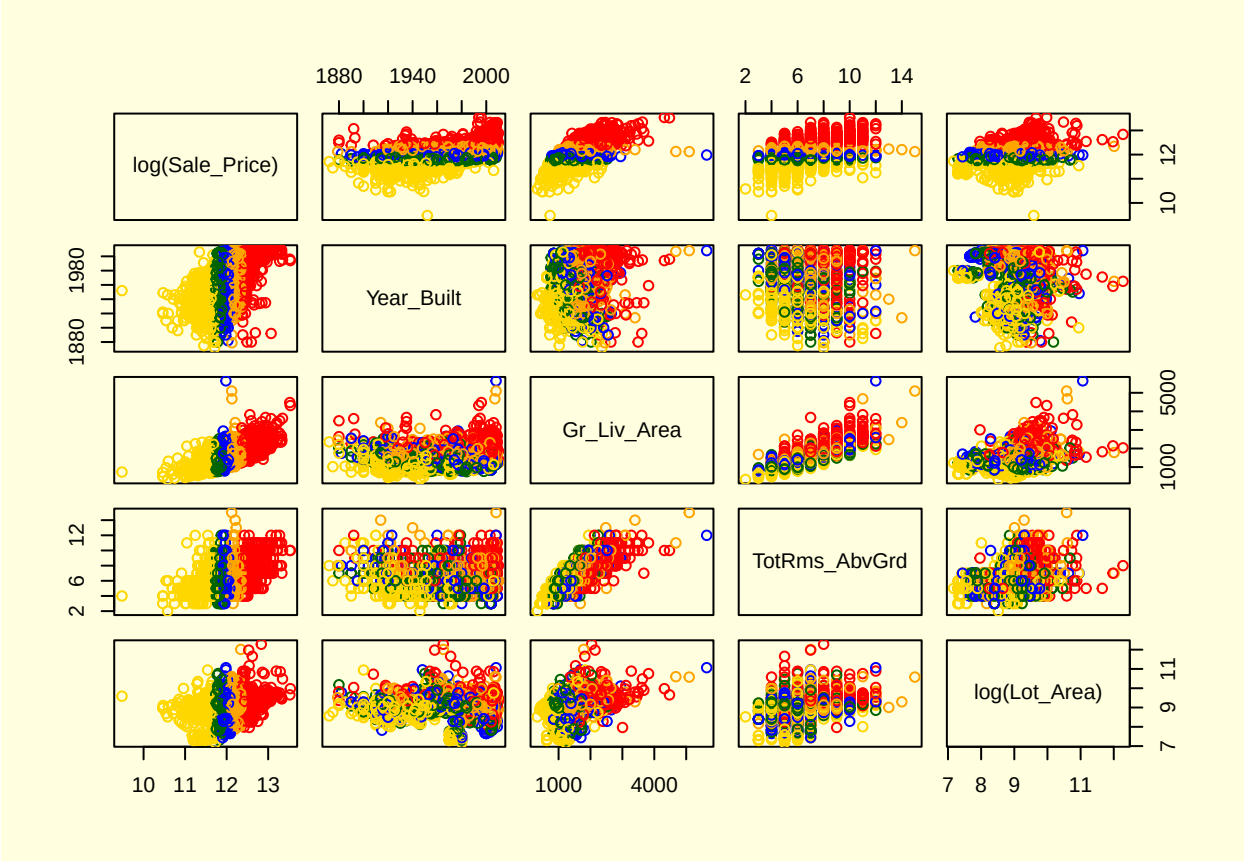

Figure 9: The scatterplot matrix of the full model with log(Sale_Price) and log(Lot_Area) and other predictors held fixed in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

The scatterplot matrix in Figure 9 shows noticeable change and improvement in the data distribution for both the response and the predictor, Lot_Area, because the data points are distributed more normally with fewer outliers than the first scatterplot matrix in Figure 7.

We rechecked the model assumptions again to ensure those initial issues had been improved.

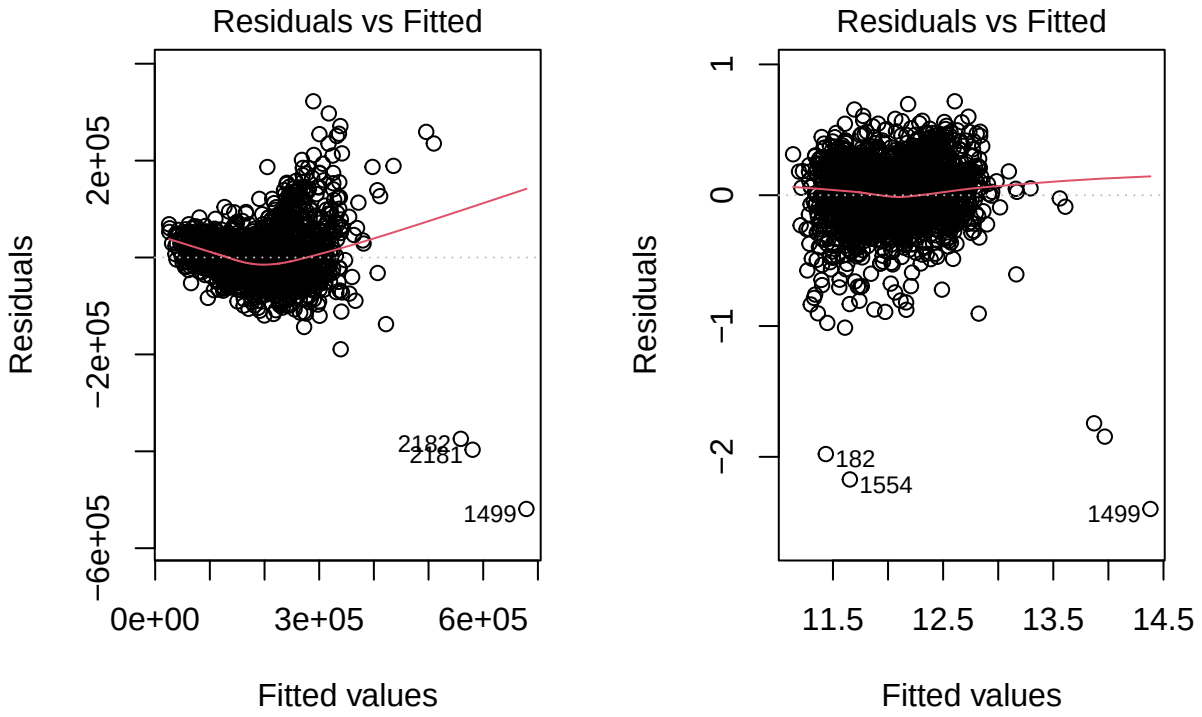

Figure 10: The residuals vs. fitted plots of the original model(L) and final model(R) with log transformation in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

The linearity of the final model on the left hand in Figure 10 is much improved compared to the original model on the right hand. The data looks more constant around the line though there are still some outliers.

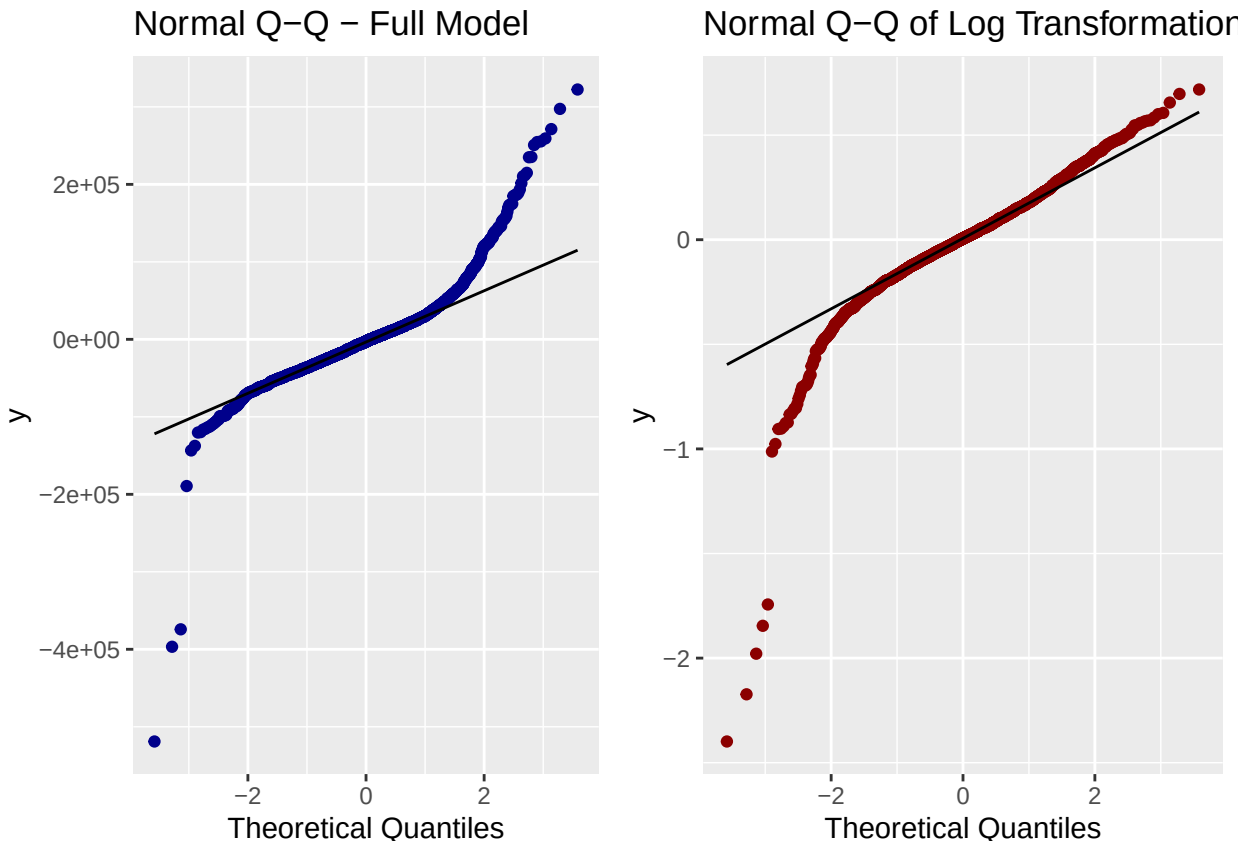

Figure 11: QQ-Plot of the original model and final model in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

The normality is also improved, as we can see on the QQ-Plot of the final model with log transformation. The data points fit better around the linear line though a heavy tail is still at the bottom. As a result, the log transformation does help reduce the skewness, linearize the relationship between the predictor and response, and reduce the impact of outliers.

Finally, we compute the full model with log transformation. Here is the final model:

$$\widehat{log(Price)} = \hat{\beta}_0 + \hat{\beta}_1(Year\_Built) + \hat{\beta}_2(Gr\_Liv\_Area) + \hat{\beta}_3(TotRms\_AbvGrd) + \hat{\beta}_4(log(Lot\_Area)$$

```
##
## Call:
## lm(formula = log(Sale_Price) ~ Year_Built + Gr_Liv_Area + TotRms_AbvGrd +
##     log(Lot_Area), data = ames3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.39857 -0.10696  0.00634  0.12039  0.71803
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.328e+00  2.825e-01  -8.238 2.61e-16 ***
## Year_Built     6.406e-03  1.372e-04  46.682  < 2e-16 ***
## Gr_Liv_Area    4.896e-04  1.404e-05  34.867  < 2e-16 ***
## TotRms_AbvGrd -2.796e-02  4.351e-03  -6.426 1.52e-10 ***
## log(Lot_Area)  1.282e-01  8.453e-03  15.170  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2153 on 2925 degrees of freedom
## Multiple R-squared:  0.7213, Adjusted R-squared:  0.7209
## F-statistic:  1892 on 4 and 2925 DF,  p-value: < 2.2e-16
```

Table 6: The summary of the original and final models with log transformation in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

The coefficients indicate the estimated effect of each predictor variable on the log-transformed sale price, holding other variables constant. For example, the coefficient for the year built (6.406e-03) suggests that, on average, a one-year increase in the year built is associated with a 0.64% increase in the sale price, holding other variables constant.

The p-values of all predictor variables are less than 0.05, indicating that they are statistically significant predictors of the log-transformed sale price.

The multiple R-squared indicates that the model explains about 72.13% of the variability in the log-transformed sale price. The adjusted R-squared value of 72.09% considers the number of predictor variables in the model. The F-statistic (1892) and associated p-value (< 2.2e-16) suggest that the model as a whole is statistically significant.

The residuals are the differences between the observed sale prices and those predicted by the model. The residual standard error (0.2153) is the average amount by which the observed sale prices differ from the predicted values and can be used to assess the model's overall accuracy.

We also made predictions about the sale price of a house by picking randoms numbers for Gr_Liv_Area of 1000 square feet, Year_Built of 1990 and 2008, TotRms_AbvGrd = 5, and Lot_Ara = 10000, and here is what we found:

```
##        fit      lwr      upr
## 1 155007.0 101596.3 236496.4
## 2 173951.7 114000.9 265429.3
```

Table 7: Predicted sale price on the natural logarithm model in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

From the results of Table7, we are 95% confident that the actual price of a single home is between $101,596.3 and $236,496 for the house that was built in 1990 and between $111,000 and $265,430 for the house that was built in 2008 with 1000 square feet above the ground living area, five rooms above the ground = 5, and 10000 square feet lot area.

Conclusion

We used linear regression to analyze data from the AmesHousing dataset in Ames, Iowa between 2006 and 2010 to find out what factors affect house prices. First, we found out that the size of the living area and lot area were found to be positively associated with the sale price. This means that the larger the living area or lot area, the more expensive the home is likely to be.

Similarly, the number of bedrooms (without bathrooms) in a home was also found to be a significant predictor of the sale price. This indicates that homes with more bedrooms tend to be more expensive.

On the other hand, the age of a home was found to have a negative impact on the sale price which makes sense that the older the house, the less expensive than the newer ones. And overall, our model was able to explain about 72% of the variation in sale prices based on these variables. However, there are some limitations with outliers and other factors like location, and school district might also influence the prediction and analysis.

In summary, our analysis found that the size of the living area and lot area, the number of bedrooms, and the age of the house are important factors in determining the sale price of a home in Iowa. We think that this information could be helpful for people both in the real estate industry and for anyone looking to buy or sell a house.

Appendices

R code:

Subset data set from AmesHousing dataset in Ames, Iowa, between 2006 and 2010 (data obtained from R package).

```
ames <- make_ames()
ames2 <- ames %>%
   select("Sale_Price", "Gr_Liv_Area", "Year_Sold",
          "TotRms_AbvGrd", "Year_Built", "Utilities",
          "Lot_Area", "Street", "Fence") %>%
   mutate(Utilities = as.integer(Utilities),
         Street = as.integer(Street),
         Fence = as.integer(Fence))
```

Figure 1: The partial regression plots of 8 predictors.

```
fit.full <- lm(Sale_Price ~ ., data = ames2)
avPlots(fit.full)
```

Table 1: The summary table of the full model with eight predictors and one response.

```
fit.full <- lm(Sale_Price ~ ., data = ames2)
summary(fit.full)
```

Table 2: Forward Stepwise Selection with AIC table computed with nine predictors.

```
mod_0 <- lm(Sale_Price ~ 1, data = na.omit(ames2))
mod_full <- lm(Sale_Price ~ ., data = ames2)
n <- nrow(ames2)

forward_AIC <- step(mod_0,
               scope = list(lower = mod_0, upper = mod_full),
               direction = "forward", trace = 0)
forward_AIC
```

Table 3: The summary table of the reduced model with five variables.

```
ames3 <- ames2 %>%
   select("Sale_Price", "Gr_Liv_Area", "Year_Built", "Lot_Area", "TotRms_AbvGrd")

fit.reduced <- lm(Sale_Price ~ Gr_Liv_Area + Year_Built +
                  Lot_Area + TotRms_AbvGrd, data = ames3)
summary(fit.reduced)
```

Figure 2: Diagnostic plots of the reduced model.

```
par(mfrow = c(2, 2))
plot(fit.reduced, which = 1:4)
```

Figure 3: Outlier and Leverage Diagnostics for Sale_Price.

```
ols_plot_resid_lev(fit.reduced)
```

Table 3: Data rows for outlier points obtained from the original dataset.

```
ames3[c(1499, 2181, 2182, 1500, 2180, 2183), ]
```

Figure 4: The histogram of the response, Sale_Price.

```
ggplot(ames3, aes(x = Sale_Price)) +
  geom_histogram(color = "black",
                 fill = "lightslateblue", bins = 50) +
  scale_x_continuous(labels = comma) +
  labs(title = "Distribution of house prices",
           x = "Price", y = "Frequency") +
  theme_minimal()
```

Figure 5: Profile Log-Likelihood plot of the response, Sale_Price.

```
boxCox(fit.reduced)
```

Figure 6: The side-by-side histogram of the response before and after the log transformation.

```
library(patchwork)
p1 <- ggplot(ames3, aes(x = Sale_Price)) +
  geom_histogram(bins = 30,
                 fill = "skyblue2", color = "white") +
  theme_light() +
  ggtitle("Histogram of Sale Price") +
  ylab("Count")

p2 <- ggplot(data = ames3, aes(x = log(Sale_Price))) +
  geom_histogram(bins = 30,
                 fill = "darkred", color = "white") +
  theme_light() +
  ggtitle("Log Transformation of Sale Price") +
  ylab("Count")
p1 + p2 + plot_layout(ncol = 2)
```

Figure 7: The scatterplot matrix of the full model with log(Sale_Price) and four predictors held fixed.

```
ames3$Price_Group <- cut(ames3$Sale_Price, quantile(ames3$Sale_Price,
         probs = seq(0, 1, 0.2)), labels = FALSE)

price_colors <- c("darkorange", "steelblue", "forestgreen", "red", "tomato")

pairs(log(Sale_Price) ~ Year_Built + Gr_Liv_Area + TotRms_AbvGrd + Lot_Area, data = ames3,
      col = price_colors[ames3$Price_Group])
```

Figure 8: The coefficient matrix plot of all reduced models between the response and predictors.

```
ggcorr(ames2, size = 3, label = TRUE, label_size = 4, label_round = 2, label_alpha = TRUE)
```

Table 4: Variance Inflation Factor (VIF) predictors.

```
lm_logfit1 = lm(log(Sale_Price) ~ Year_Built + Gr_Liv_Area +
              TotRms_AbvGrd + Lot_Area, data = ames3)
round(vif(lm_logfit1), 2)
```

Table 5: AIC table of two models with and without log(Lot_Area).

```
lm_logfit1 = lm(log(Sale_Price) ~ Year_Built + Gr_Liv_Area +
             TotRms_AbvGrd + Lot_Area, data = ames3)
lm_logfit2 = lm(log(Sale_Price) ~ Year_Built + Gr_Liv_Area +
              TotRms_AbvGrd + log(Lot_Area), data = ames3)
round(AIC(lm_logfit1, lm_logfit2), 2)
```

Figure 9: The scatterplot matrix of the full model with log(Sale_Price) and log(Lot_Area) and other predictors held fixed.

```
ames3$Price_Group <- cut(ames3$Sale_Price, quantile(ames3$Sale_Price,
         probs = seq(0, 1, 0.2)), labels = FALSE)
par(bg = "lightyellow")
price_colors <- c("gold", "darkgreen", "blue", "orange", "red")

pairs(log(Sale_Price) ~ Year_Built + Gr_Liv_Area +
         TotRms_AbvGrd + log(Lot_Area), data = ames3,
      col = price_colors[ames3$Price_Group])
```

Figure 10: The residuals vs. fitted plots of the original model(L) and final model(R) with log transformation.

```
par(mfrow = c(1, 2))
plot(fit.reduced, which = 1:1)
plot(lm_logfit2, which = 1:1)
```

Figure 11: QQ-Plot of the original model and final model.

```
qq1 <- ggplot(data.frame(qqnorm(resid(fit.reduced))),
              aes(sample = resid(fit.reduced))) +
  stat_qq(color = "darkblue") +
  stat_qq_line() +
  xlab("Theoretical Quantiles") +
   ylab("Residuals")+
  ggtitle("Original Model")

qq2 <- ggplot(data.frame(qqnorm(resid(lm_logfit2))),
              aes(sample = resid(lm_logfit2))) +
  stat_qq(color = "darkred") +
  stat_qq_line() +
  xlab("Theoretical Quantiles")+
  ylab("Residuals")+
  ggtitle("Final Model (with log trasf)")

plot_grid(qq1, qq2, ncol = 2)
```

Table 6: The summary of the original and final models with log transformation.

```
lm_logfit2 = lm(log(Sale_Price) ~ Year_Built + Gr_Liv_Area +
             TotRms_AbvGrd + log(Lot_Area), data = ames3)
summary(lm_logfit2)
```

Table 7: Predicted sale price on the natural logarithm model.

```
new_x <- data.frame(Gr_Liv_Area = 1000, Year_Built =
         c(1990, 2008), TotRms_AbvGrd = 5, Lot_Area = 10000)
exp(predict(lm_logfit2, newdata = new_x, type = "response", interval = "prediction"))
```

Reference

R Core Team (2021). datasets: Datasets for 'R'. R package version 4.1.0. https://CRAN.R-project.org/package=datasets

Dean De Cock (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. Journal of Statistics Education, 19(3). https://ww2.amstat.org/publications/jse/v19n3/decock.pdf