*Explores the Ames housing dataset using R with a correlation matrix, multiple regression, and residuals*

**Assignment**

**1**

A1

ALY6015 Intermediate Analytics

Assignment 1 – Regression Diagnostics

**PREPERATION:**

By: John DiSessa

For: Professor Goulding

On: September 26th,2021

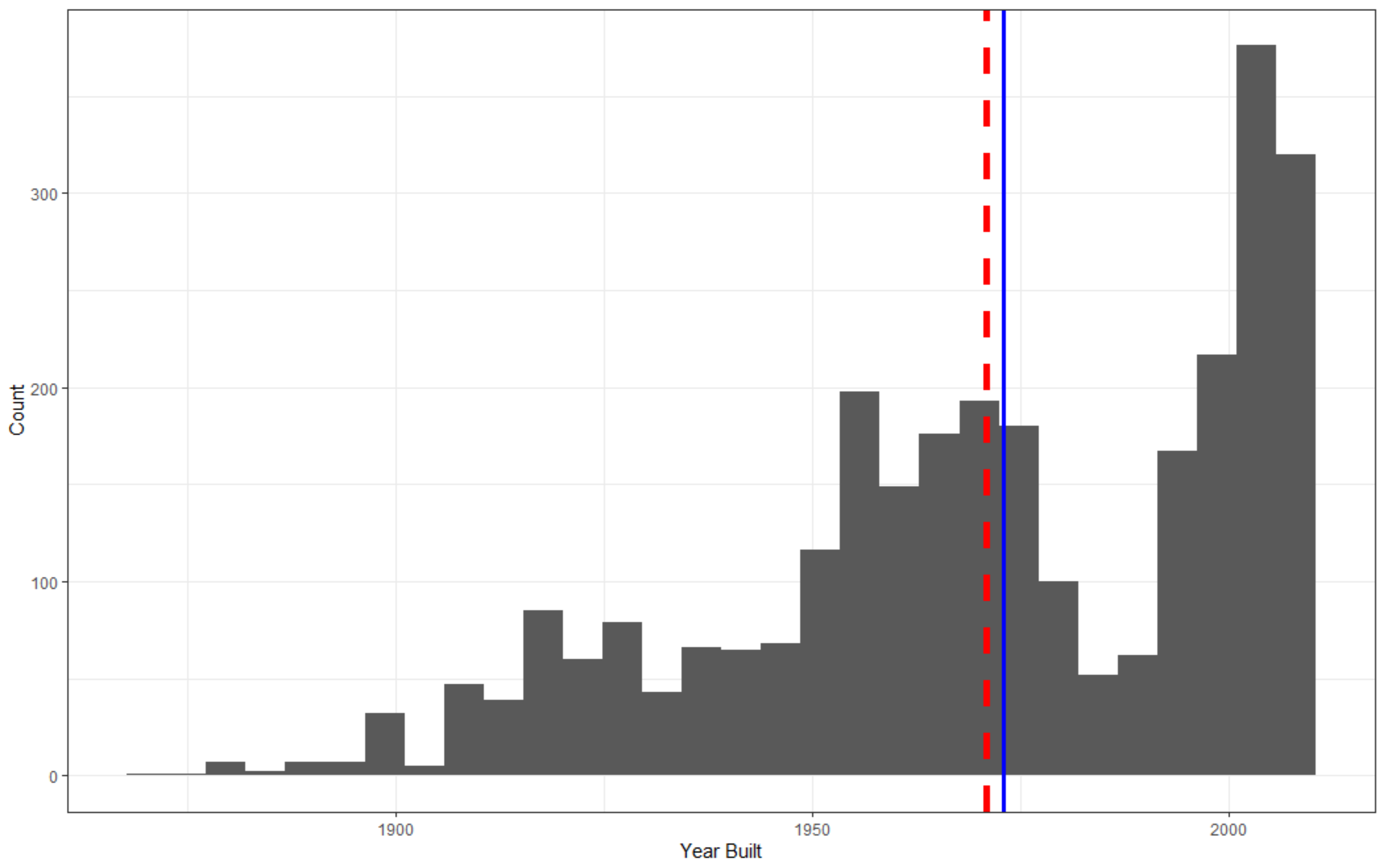
**Introduction**

I analyzed housing data from Ames, IA in order to consider the effects certain factors have on the prices of homes in the city. Using the AmesHousing.csv dataset provided, I created histograms, scatter plots, and regression models to summarize the data and identify significant predictors of Sale Price for homes in Ames.

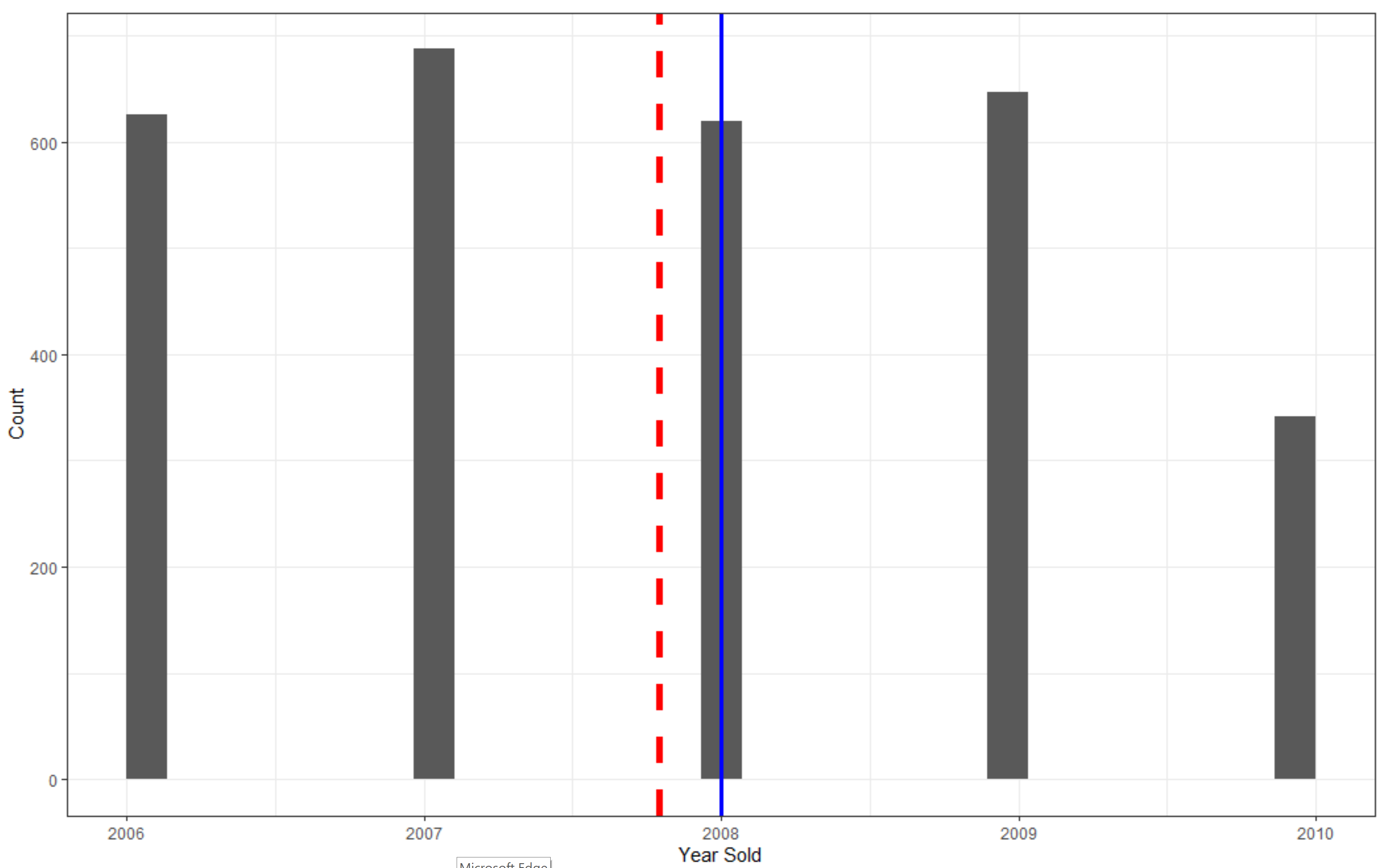
**Exploratory Data Analysis**

First, I removed several outliers from the dataset that were not representative of the Ames, IA housing market. A few homes had excessive square footage data points for 1st floor area or lot area. Second, I removed entries that had year built or garage year built that were incorrect. There was one example of a home built in 2207, which clearly hasn’t happened yet. After I removed outliers, I found the average home by taking the means of different variables. For example, the average home in our dataset was built in 1971 and sold in 2007 for $180,223. I then calculated the coefficient of variation to see if our Sale Price variable is normal or excessive. With a CV = .43 Sale Price is not too variable and we can continue working with this normal dataset.

Next, I created a histogram to see the distribution of what years our homes were built. The mean year of 1971 is marked by the dotted red line. The median year of 1973 is marked by the solid blue line. The Year Built variable is slightly left skewed since the mean is less than the median and this is proven by seeing all of the incredibly early years of home building.



I then conducted similar analysis with the Year Sold variable since knowing what years our houses were sold could directly affect the sale price given inflation, cost of living changes, and the 2008 recession. The mean is slightly less than the median (Sept 2007 < June 2008) but these values are very similar so it is safe to assume an even distribution.

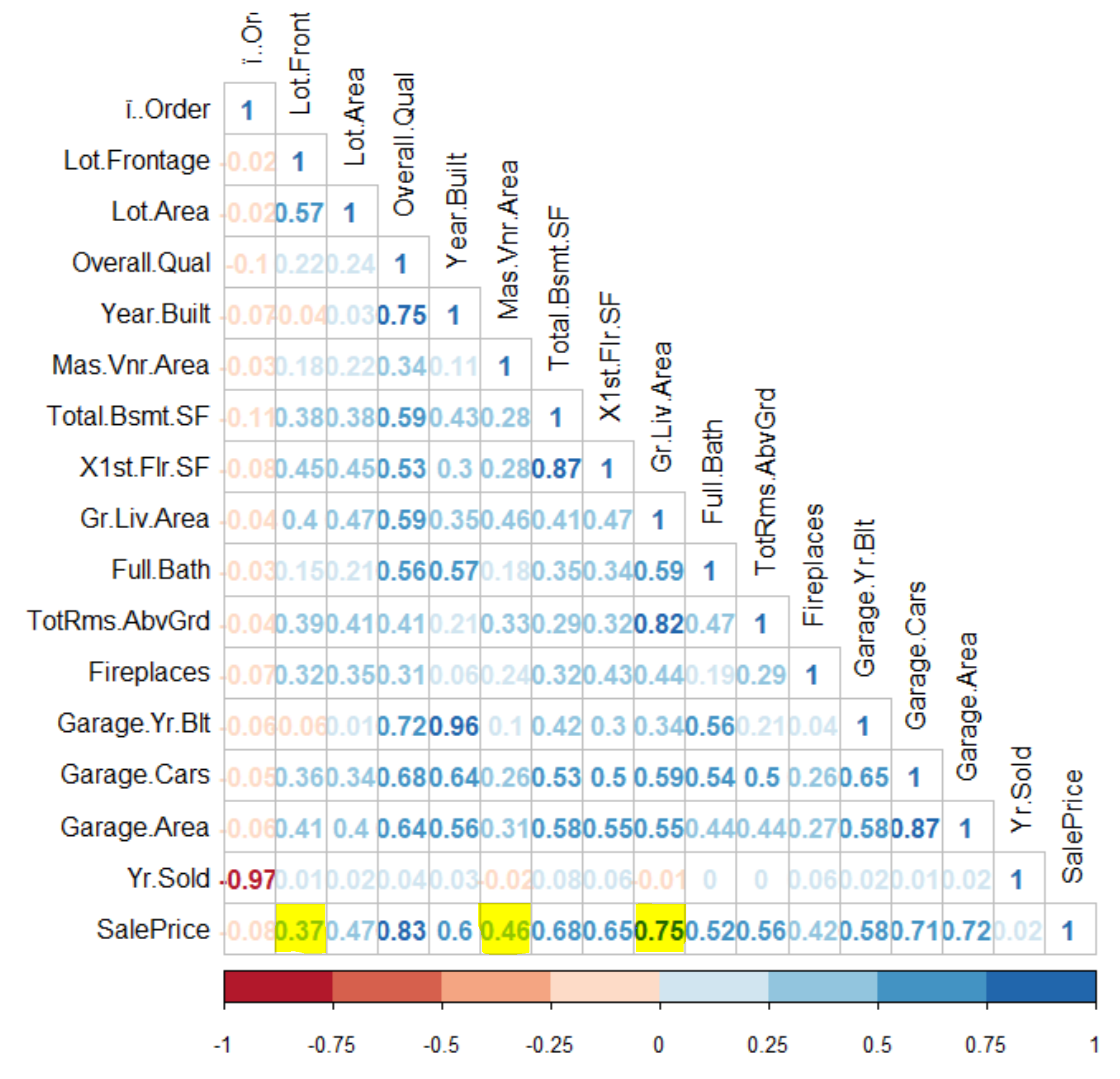


Lastly, I created a scatter plot to see how Sale Price was influenced by the year a house was built in order to estimate how much better and more valuable newer homes really are. Starting around 1990 there is an inflection point where Sale Price really starts to increase at a faster rate than previously seen. Fewer and fewer Sale Prices fell below the mean Sale Price line of $180,223.



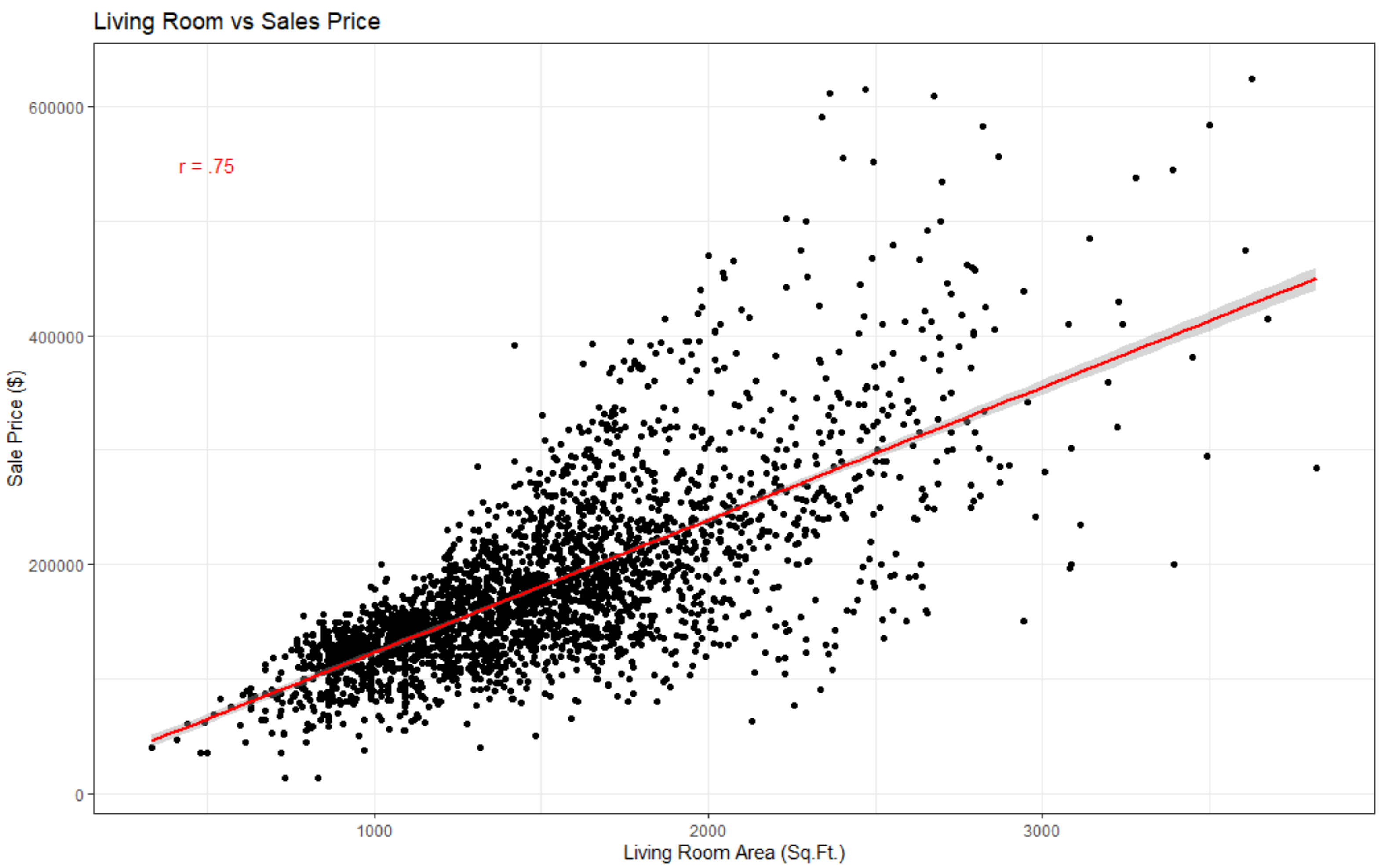
**Correlation**

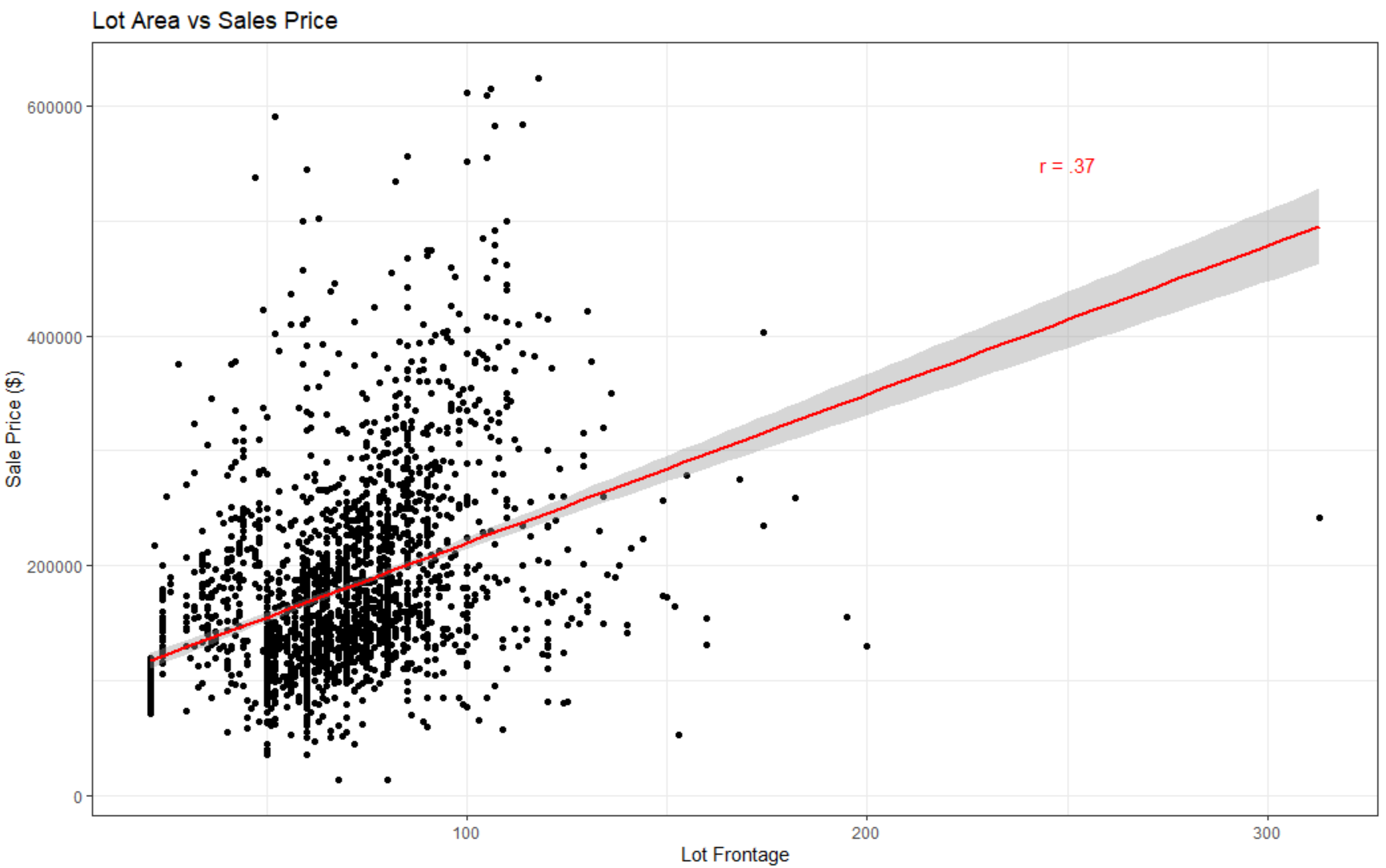
Besides looking at a few variables I previously explored, I needed to look at all of the variables and see how they affect a home’s sale price. I created the following correlation matrix to see all of the variable’s correlations with Sale Price. This chart also helped me potentially identify multicollinearity if any dependent variables were correlated with each other.

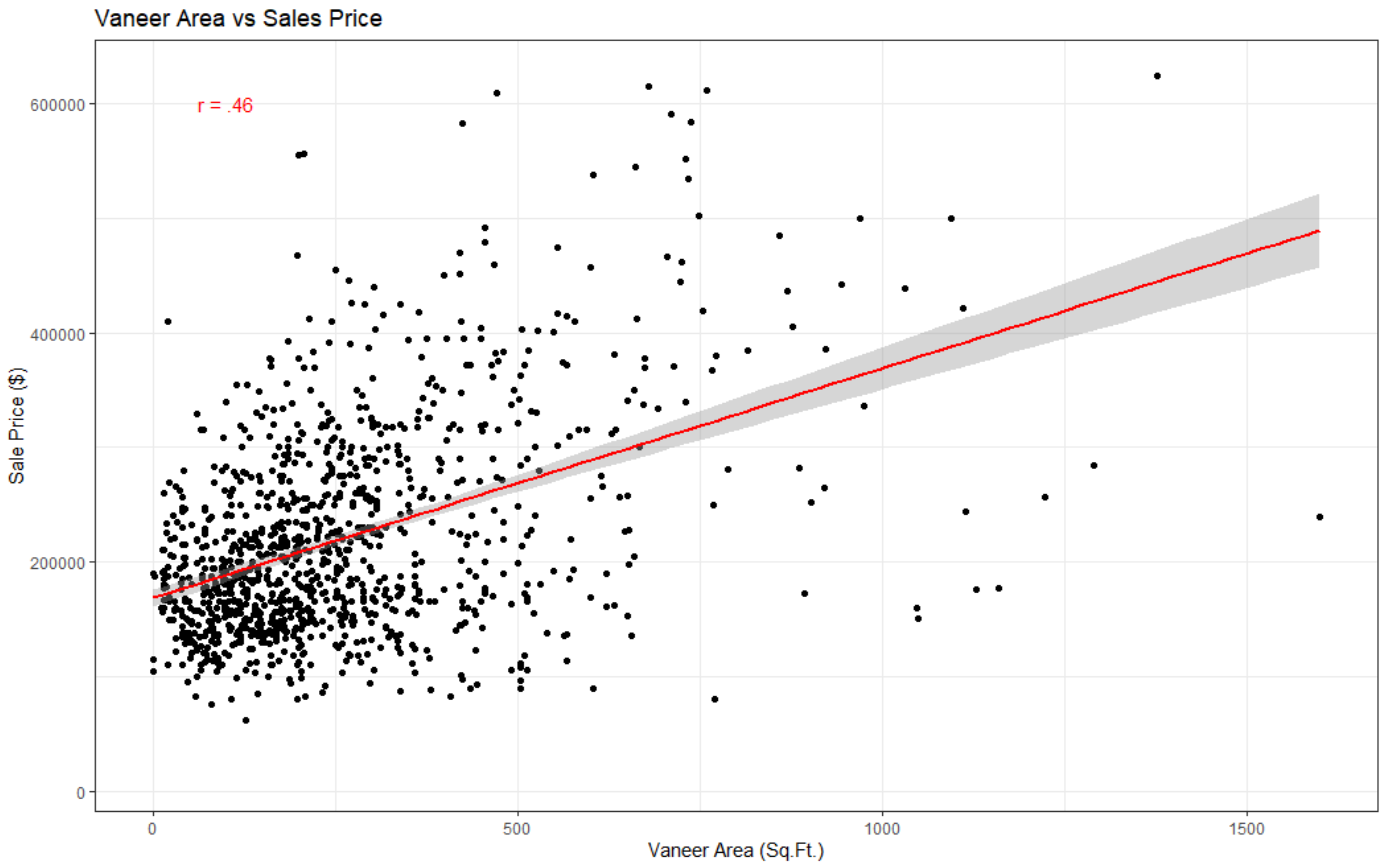


Normally I would not publish a correlation matrix with all these r values since it is difficult to read. However, it was critical for identifying the highest, lowest, and average variables correlated with Sale Price as seen with the yellow highlights. Before analyzing these three correlations, I ran a p-value correlation matrix and confirmed than they all had p-values less than my alpha of .05.

The scatter plots below show the very strong positive correlation between Living Room Area and Sale Price, the weak positive correlation between Lot Frontage Area and Sale Price, and a medium positive correlation between Masonry Vaneer Area and Sale Price. Their respective r values, as shown by the red annotations on each graph, are .75, .37, and .46. Even though they look similar from a quick glance, it is misleading since the scale of each x axis value is different. Additionally, there were no variables negatively correlated with Sale Price which makes sense considering adding more fireplaces or square footage to a house would never make it less valuable. What is more interesting, however, is the respective confidence interval as seen in they gray areas around the mean lines. The strongest correlated variable has the smallest confidence interval and the weakest correlated variable has the largest confidence interval.

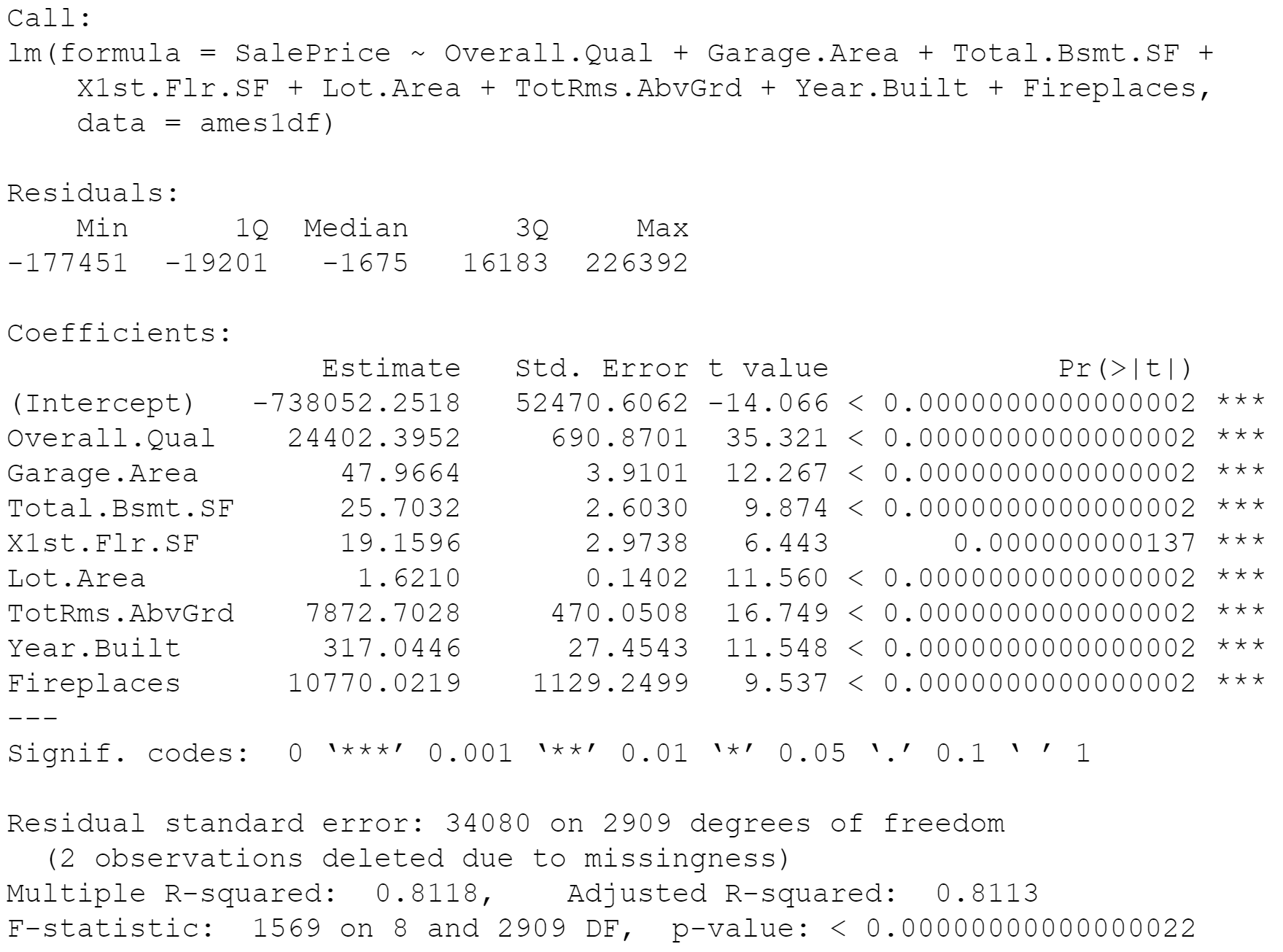






**Regression**

In order to predict a Sale Price given certain variables, I created a multiple linear regression model. The first model I created had some variables that were only marginally significant. It also contained multicollinearity between Garage Area and Garage Cars. Their VIF values were over 5 and their correlation was .87. Even though that model had a high adjusted R2 value and was statistically significant, the multicollinearity was significant enough that I re-worked my model with fewer variables but a more accurate model, as seen below.



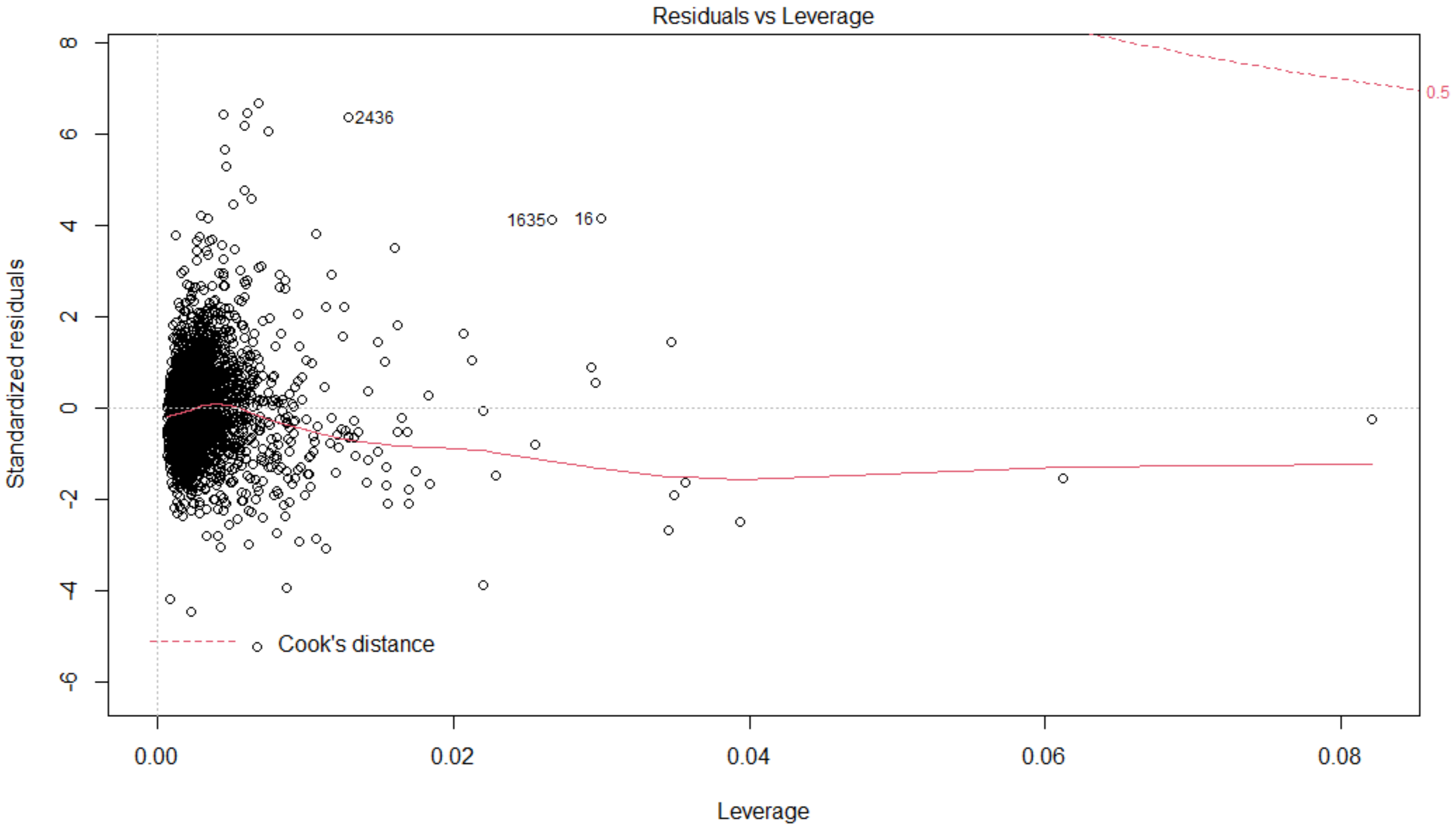
Based on these regression calculations, the following formula predicts Sale Price given these 8 variables.

*SalePrice = -738052.25 + 24402.40\*(****Overall.Qual****) + 47.97\*(****Garage.Area****) + 25.70\*(****Total.Bsmt.SF****) + 19.16\*(****X1st.Flr.SF****) + 1.62\*(****Lot.Area****) + 7872.70\*(****TotRms.AbvGrd****) + 317.04 \*(****Year.Built****) + 10770.02\*(****Fireplace****s) ± 54797.84*

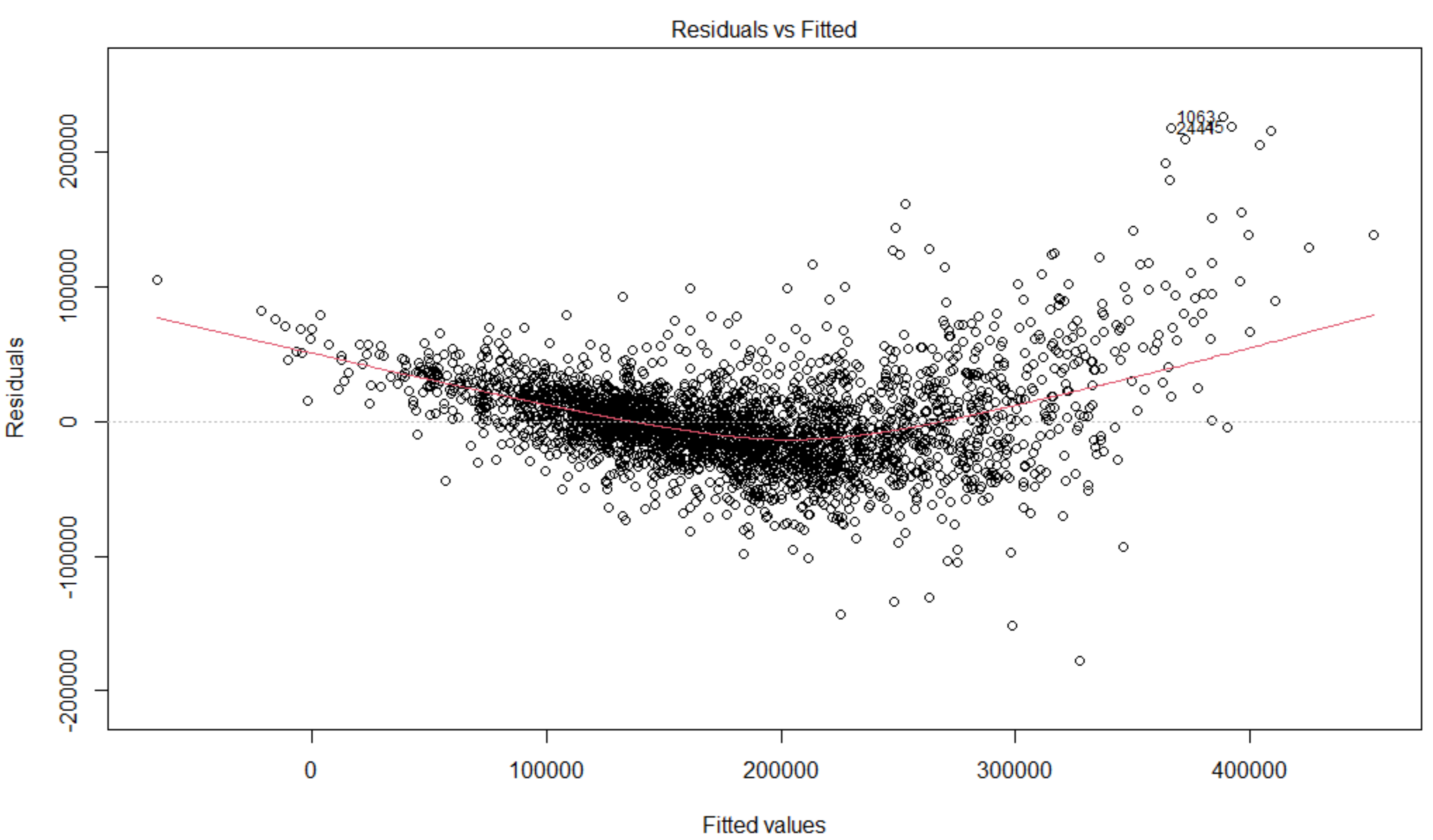
Each coefficient represents how much value a 1 unit increase in that variable adds to the sale price, holding all other variables constant. For example, a home with a 7 rating, 300 sqft garage, 500 sqft basement, 1000 sqft first floor, 10000 sqft lot area, 8 rooms above ground, built in 1980, and has 1 fireplace is worth on average $251,660. If I added 1 room above ground to the house, holding all other variables constant, the home would then be worth an additional $7,872.

Since this model is statistically significant (p-value < .05) and the adjusted R2 is .8113, we can say this is an accurate model and that 81.13% of the variation in Sale Price can be explained by these 8 variables.

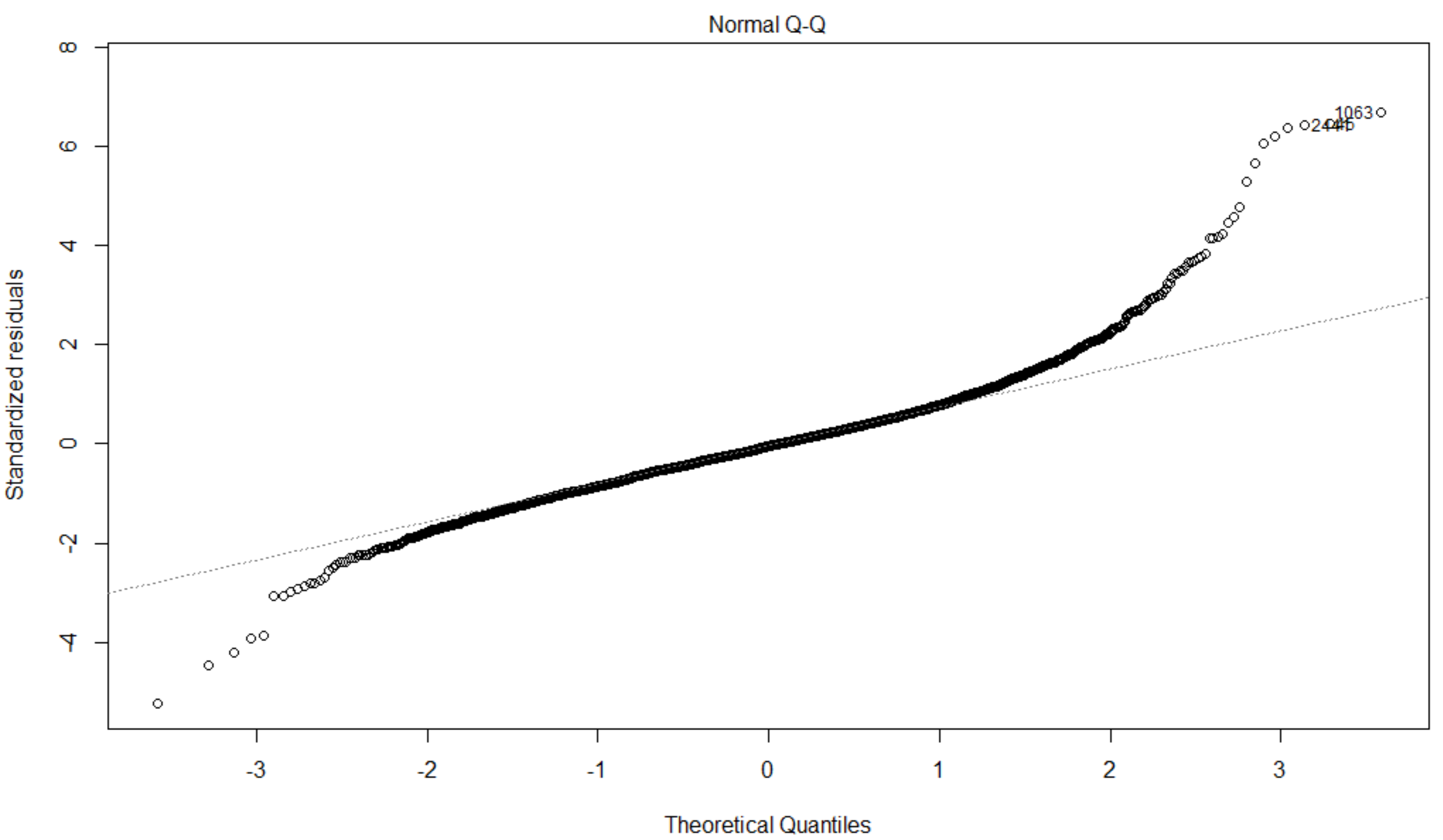
However, just because this model is a great predictor of Sale Price, it does not necessarily mean that it fits the data that well. In order to see if it does fit the data well, I first created a Residuals vs Leverage plot. This chart shows us if there are any outliers or significantly influential points that, if we removed them, would alter the regression. Since no points fall out of the .5 Cook’s distance, there are no data points that are influential enough that need to be removed.



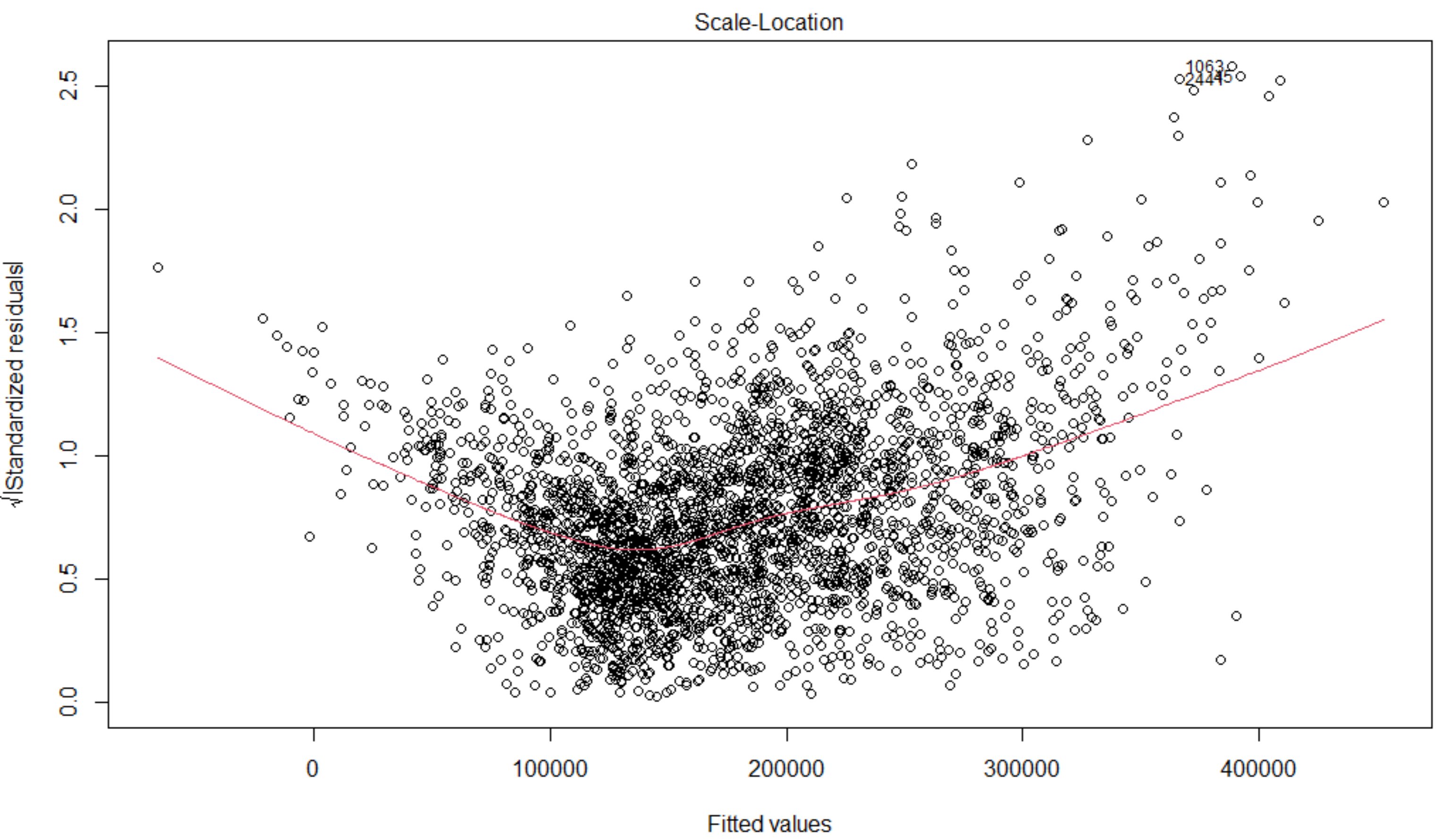
Next I created a Residuals vs Fitted graph to show if there is a linear or non-linear relationship between our model’s variables and Sale Price. Since we created a multiple *linear* regression model, we assume that small values and large values are equally predicted. However the chart below shows this is not the case. It is slightly parabolic so that values on both ends are less accurately predicted than values near the average. However, it is not extremely parabolic so it is hard to say definitively that this model is not accurate.



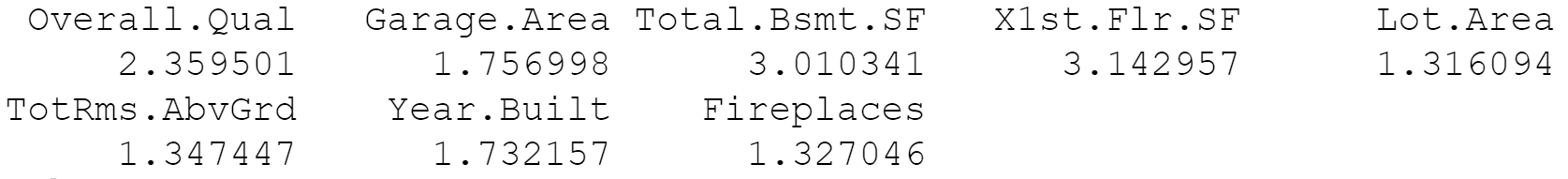
I then created a Q-Q plot to see if Sale Prices from our sample dataset come from the whole Ames, IA population that is normally distributed. Most values fall along the line so it looks like it is fairly normally distributed, however the values on both tails indicate that our dataset contains more values near the average than what is truly represented from the population data. That is not that extreme so we certainly can trust our sample dataset.



Lastly, I created a Scale-Location plot to see if the ranges of values are spread equally across our predictor variables. When I started my analysis I assumed the dataset generally had equal variance. Rather than checking the variance for each variable one at a time, I created this graph which actually shows heteroskedasticity. There is unequal variability across our predictor variables.

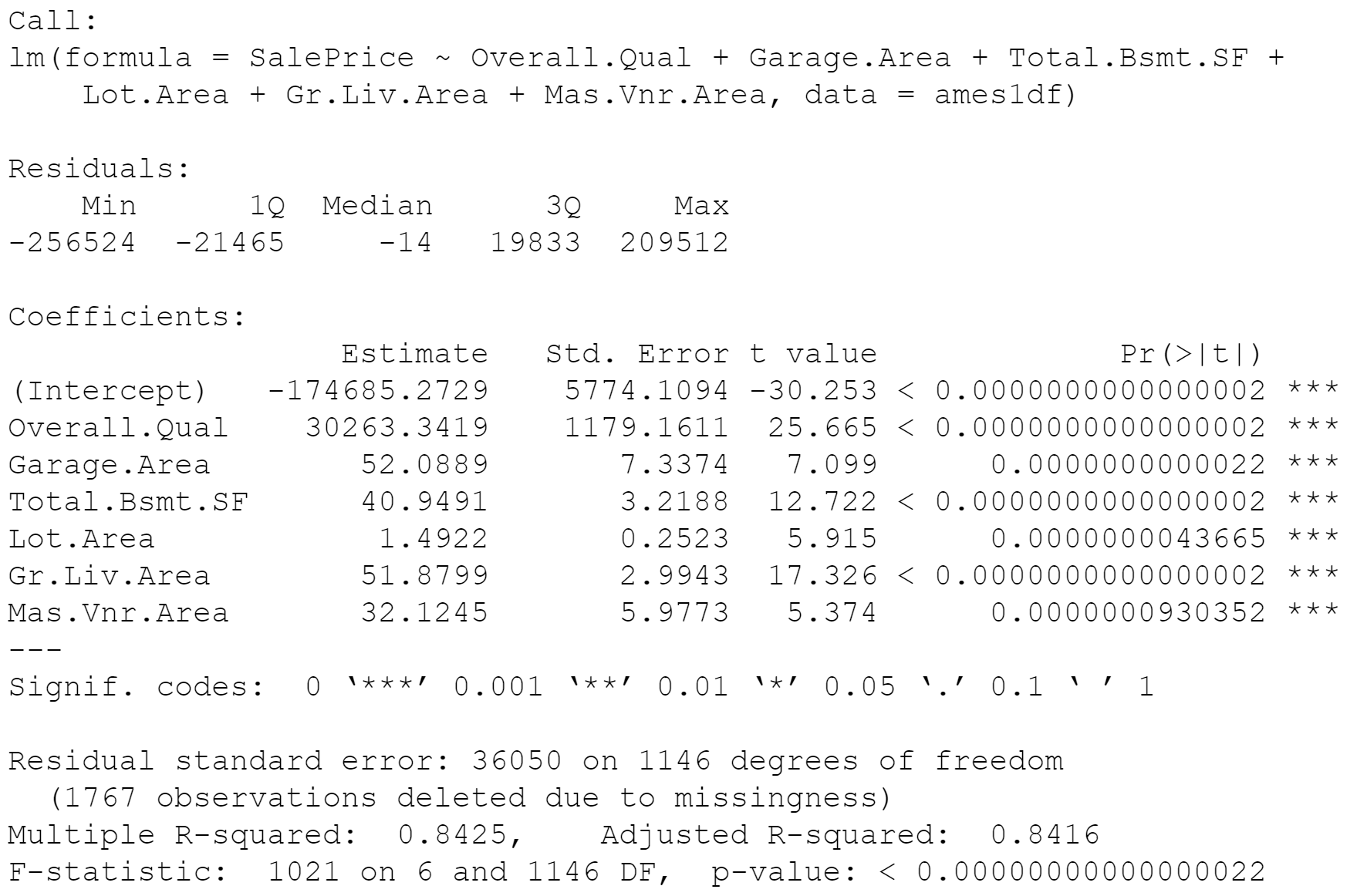


Since the graphs above did not indicate any serious problems with our regression model, I had to check again for multicollinearity. Since none of these predictors are correlated with each other (Variance Inflation Factor < 5) we can check the last box and finally confirm that my model is a good predictor of Sale Prices for homes in Ames, IA.



**Better Model**

After all that work in confirming my model is good, I wanted to see if there was actually a great model or if mine was the best possible model given the dataset. I used the submodels function in R to see the best number of predictors to include in the model and which predictors to use in the new model. After guessing and checking, I found how many and which variables I should include for the highest adjusted R2, the lowest CP values (standardized variation), and the lowest BIC values (test error). A model with 8 predictors was the highest amount of predictors that all 3 metrics agreed on was the best. However, with 8 and 7 predictors, the best model used both Garage Area and Number of Cars. Since I already previously identified the multicollinearity between them, I settled on the model with the most predictors that only included one of these values. 6 was our magic number.



*SalePrice = -174685 + 30263\*(****Overall.Qual****) + 52\*(****Garage.Area****) + 41\*(****Total.Bsmt.SF****) + 1.50\*(****Lot.Area****) + 52\*(****Gr.Liv.Area****) + 32\*(****Mas.Vnr.Area****) ± 6972.75*

My new model seems to be more accurate on all accounts. It has a higher adjusted R2 value, still statistically significant, a lower standard error, no multicollinearity, and the 4 ‘plot()’ function graphs indicate slightly more linear fit, slightly less variation, and still no outliers that significantly impact the model. I am confident in my new model being the best possible model I could have created given the dataset.

**Summary**

After basic exploratory data analysis, identifying important correlations, building two different multiple linear regression models, and verifying their fit, accuracy, and volatility, I created the most accurate model to predict the selling prices of homes in Ames, IA. I used histograms, scatterplots, correlation matrices, regression outputs, leverage graphs, fitted graphs, Q-Q plots, and scale-location plots to demonstrate my knowledge of statistics and R for our first assignment of the course.

**Citations**

Bevans, Rebecca. “Multiple Linear Regression: A Quick and Simple Guide.” *Scribbr*, 26 Oct. 2020, www.scribbr.com/statistics/multiple-linear-regression/.

Varshney, Paras. “Q-Q Plots Explained.” *Medium*, Towards Data Science, 17 Oct. 2020, towardsdatascience.com/q-q-plots-explained-5aa8495426c0.

“QQ-Plots: Quantile-Quantile Plots - R Base Graphs.” *STHDA*, www.sthda.com/english/wiki/qq-plots-quantile-quantile-plots-r-base-graphs.

Stephanie. “Multicollinearity: Definition, Causes, Examples.” *Statistics How To*, 16 Sept. 2020, www.statisticshowto.com/multicollinearity/.

Holtz, Yan. “Linear Model and Confidence Interval in ggplot2.” *– The R Graph Gallery*, https://www.r-graph-gallery.com/50-51-52-scatter-plot-with-ggplot2.html.

Kuhn, Max. “The Ames IOWA Housing Data [R Package AmesHousing Version 0.0.4].” *The Comprehensive R Archive Network*, Comprehensive R Archive Network (CRAN), 23 June 2020, https://cran.r-project.org/web/packages/AmesHousing/.

Kassambara, and Sfd. “Best Subsets Regression Essentials in r.” *STHDA*, 11 Mar. 2018, http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-regression-essentials-in-r/.

Zach. “How to Plot Multiple Linear Regression Results in r.” *Statology*, 23 Dec. 2020, https://www.statology.org/plot-multiple-linear-regression-in-r/.

Lathiya, Krunal. “Variance in r: How to Use Var() Function in r.” *R*, 22 Sept. 2021, https://r-lang.com/how-to-calculate-variance-in-r-using-var-function/.

Zach. “What Is a Residuals vs. Leverage Plot? (Definition & Example).” *Statology*, 7 Sept. 2021, https://www.statology.org/residuals-vs-leverage-plot/.

Bommae, Written by. “University of VIRGINIA Library Research Data Services + Sciences.” *Research Data Services + Sciences*, https://data.library.virginia.edu/diagnostic-plots/.

Varshney, Paras. “Q-q Plots Explained.” *Medium*, Towards Data Science, 17 Oct. 2020, https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0.

jcf2d, Written by. “University of VIRGINIA Library Research Data Services + Sciences.” *Research Data Services + Sciences*, https://data.library.virginia.edu/understanding-q-q-plots/.

*The Scale Location Plot: Interpretation in r - Boostedml*. https://boostedml.com/2019/03/linear-regression-plots-scale-location-plot.html.

Choksi, Yash. “Model Selection: Cp, AIC, BIC and Adjusted R2.” *Medium*, Analytics Vidhya, 9 Mar. 2020, https://medium.com/analytics-vidhya/model-selection-cp-aic-bic-and-adjusted-r2-6a0af25945b6.

“11.3 - Best Subsets REGRESSION, Adjusted R-SQ, Mallows Cp.” *11.3 - Best Subsets Regression, Adjusted R-Sq, Mallows Cp | STAT 462*, https://online.stat.psu.edu/stat462/node/197/.