Final Project - BUSN 3040

Spencer Katzman

2024-10-07

Background & Descriptive Statistics

Context and Research Question

The analysis found in this report relates to houses sold in various Iowa neighborhoods between 2006 and 2010. I will use the data provided to run a linear regression model. This model will allow me to predict the selling price of a house in Iowa based on various physical characteristics (e.g. square footage and whether it has a finished basement). After identifying the individual impacts of those characteristics on selling price, the results can be used in several useful ways. For example, the increase in value from adding living space or refinishing a basement could be calculated, thus making the decision to do so more informed. Or, a real estate agent could use the model results to better predict the selling price of a new listing by inputting the characteristics of the house into the model.

Data and Variables - Descriptive Statistics

My initial regression model leverages the 1,460 house sales captured in the data set and uses the selling price of a house as the dependent variable (later, a revised version of the model is offered where 89 outliers are removed). Of the 80 columns of data regarding house characteristics, I have selected the square footage of the house, the size of the lot on which the house is built, an overall condition rating, and the number of bedrooms as my first four independent variables. In addition, I used data on the size of the a house's finished basement to create a dichotomous, categorical variable that represents whether or not a house has a finished basement (this variable is assigned "1" if there is a finished basement and a "0" if there is not). It is included as a fifth independent variable. I included each of these independent variables because I expect that an increase in each, other things being equal, should cause a house to sell for a higher price. My model will allow me to test these hypotheses.

A closer examination of the variables show a mean (standard deviation) for *Selling_Price* of $180,921.20 (79442.5). *Square_Footage* of livable area above grade is a quantitative continuous variable with mean (standard deviation) of 1515.464 (525.5). *Lot_Size* is quantitative continuous variable measured in square feet with mean (standard deviation) 10516.83 (9981.3). *Num_Bedrooms* is represented by a quantitative discrete variable with mean (standard deviation) 2.8664(0.815778). *Cond_Rating* is an ordinal qualitative discrete variable with mean (standard deviation) 5.5753 (1.112799). Finally, *Fin_Basement* is a qualitative variable with discrete values of 0 and 1, its mean (standard deviation) is

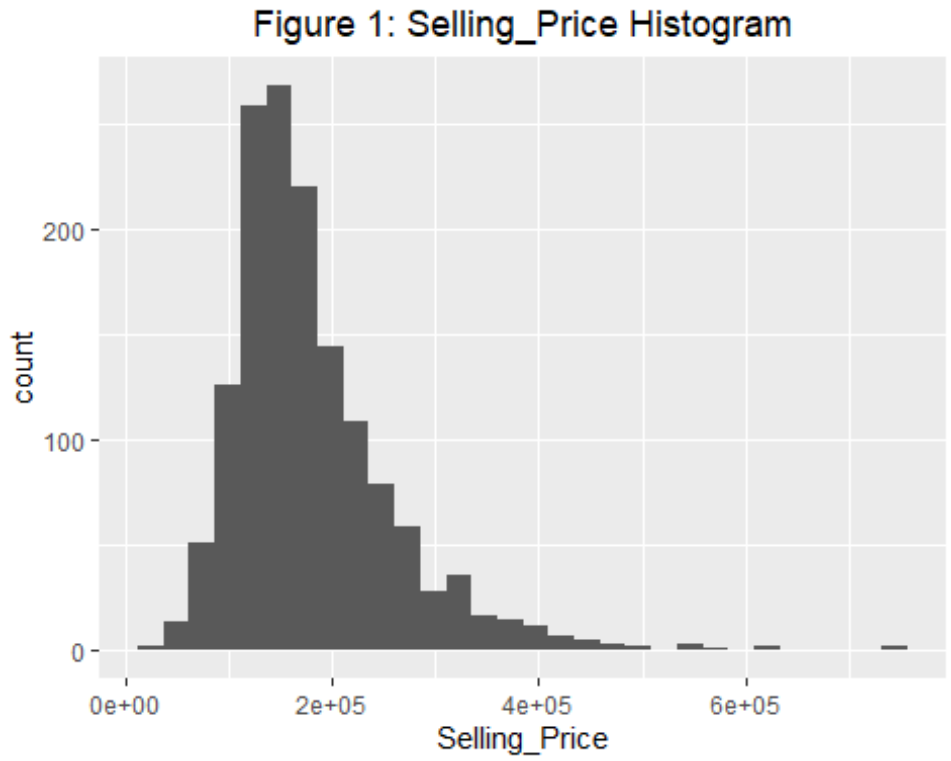
0.114 (0.3183856). This means that approximately 11.4% of the houses in the data have a finished basement.

Discriptive Visualizations

The following subsections provide visual descriptions of the dependent and independent variables in isolation.

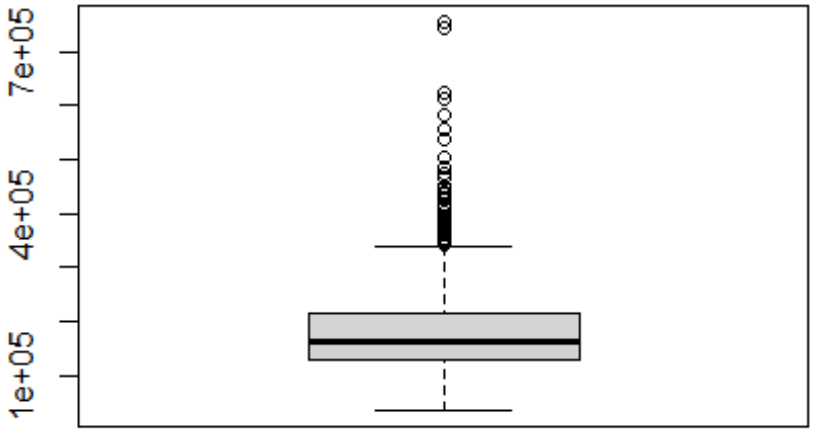
Sale Price

The following graph shows a histogram for the *Selling_Price* in the data.



The histogram in Figure 1 shows that most observations are grouped around the mean, but that there are a number of houses that sold for quite a bit more. This suggests the possible presence of outliers. The box plot in Figure 2 confirms this and I will discuss these outliers later in this report.

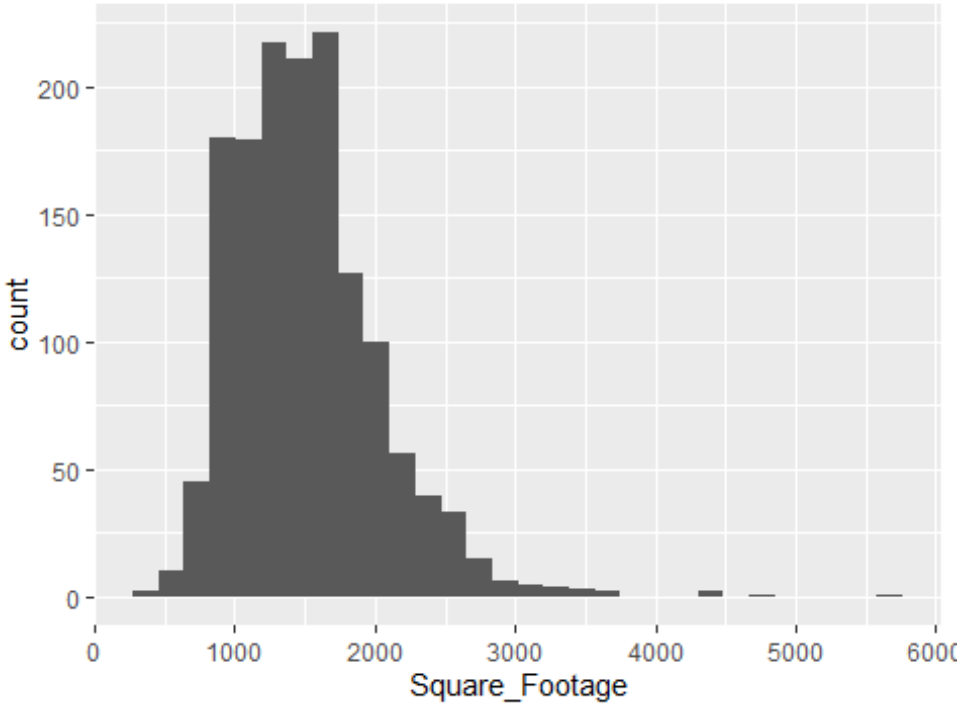
Figure 2: Selling_Price Box Plot



Square Footage

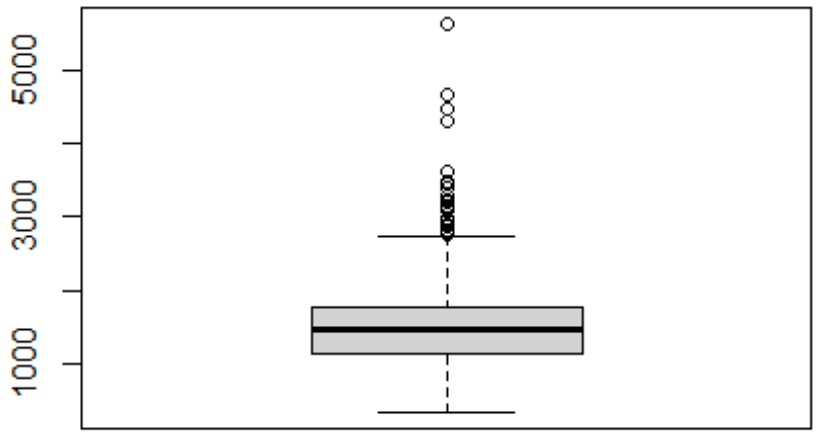
Here is the histogram for the *Square_Footage* data.

Figure 3: Square_Footage Histogram



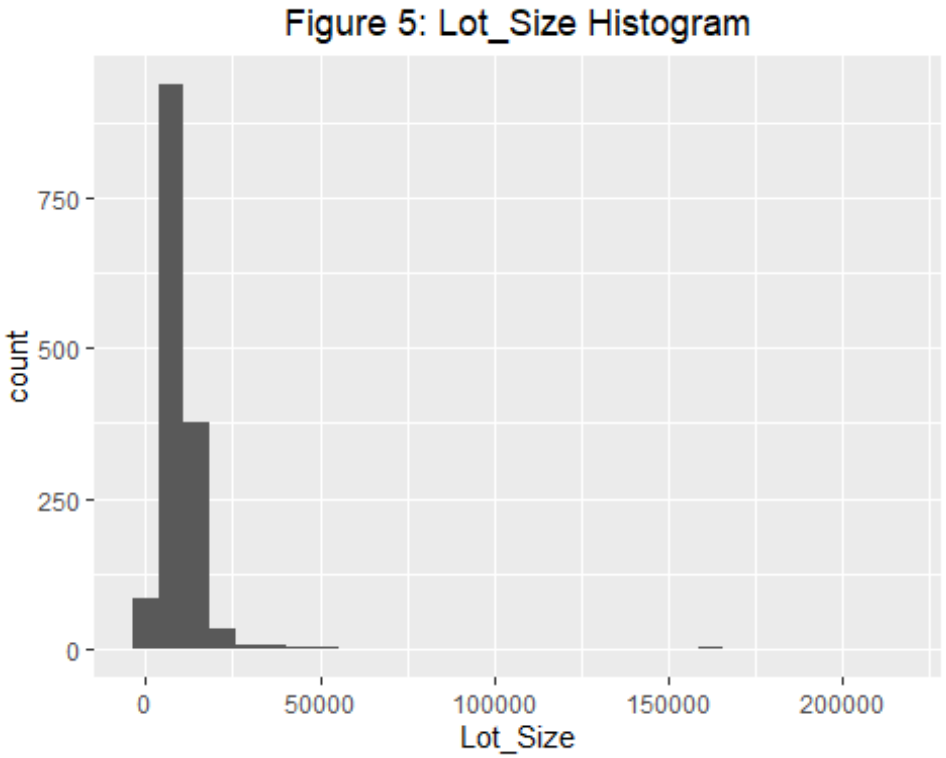
Once again, a few observations in the histogram are further to the right than most of the others, again suggesting the possible existence of outliers. The box plot in Figure 4 confirms this.

Figure 4: Square_Footage Box Plot



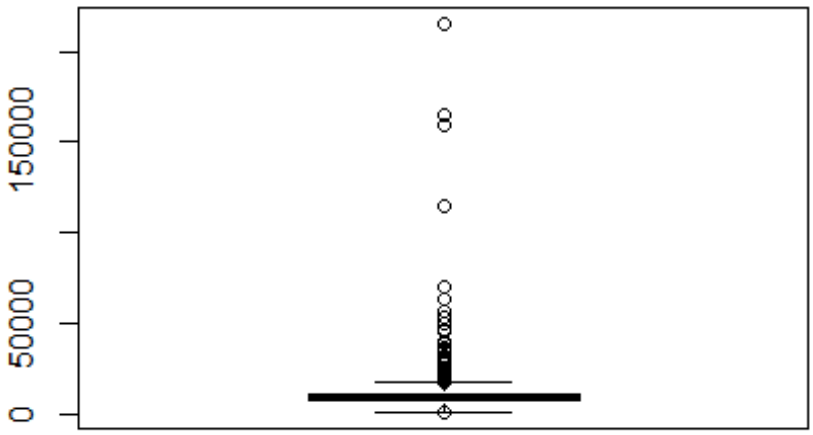
Lot Area

Figure 5 is the histogram for the size of the lot on which the house is built.



As with *Square_Footage*, the histogram for *Lot_Size* suggests the presence of large outliers. Once again, I turn to a box plot (Figure 6) to examine this possibility.

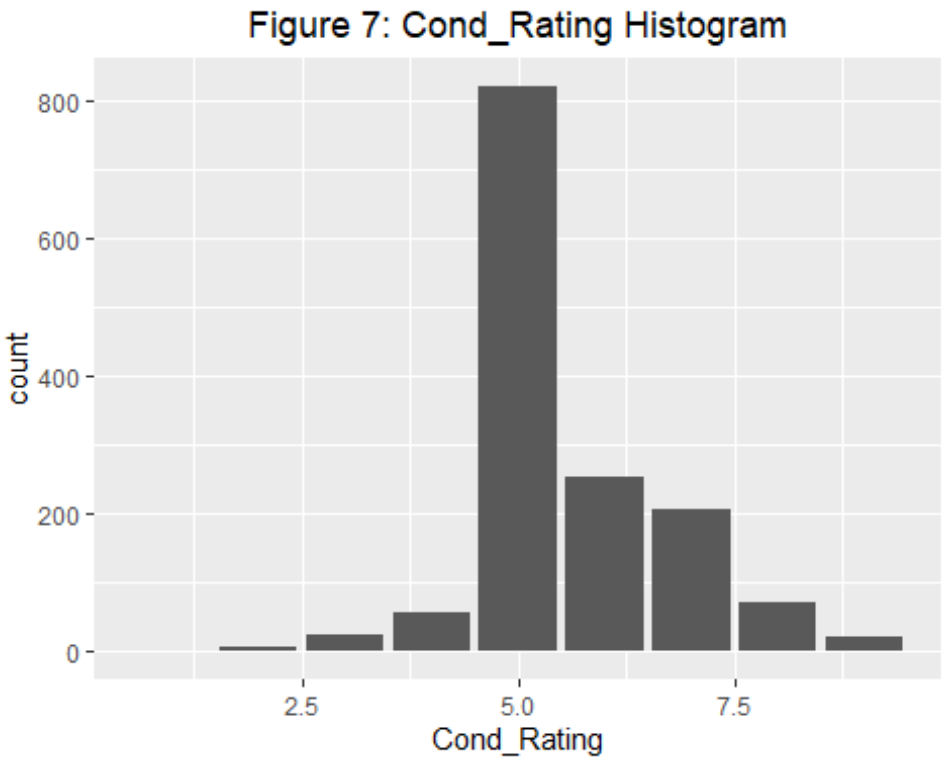
Figure 6: Lot_Size Box Plot



Indeed, the box plot verified our suspicion of large outliers for *Lot_Size*.

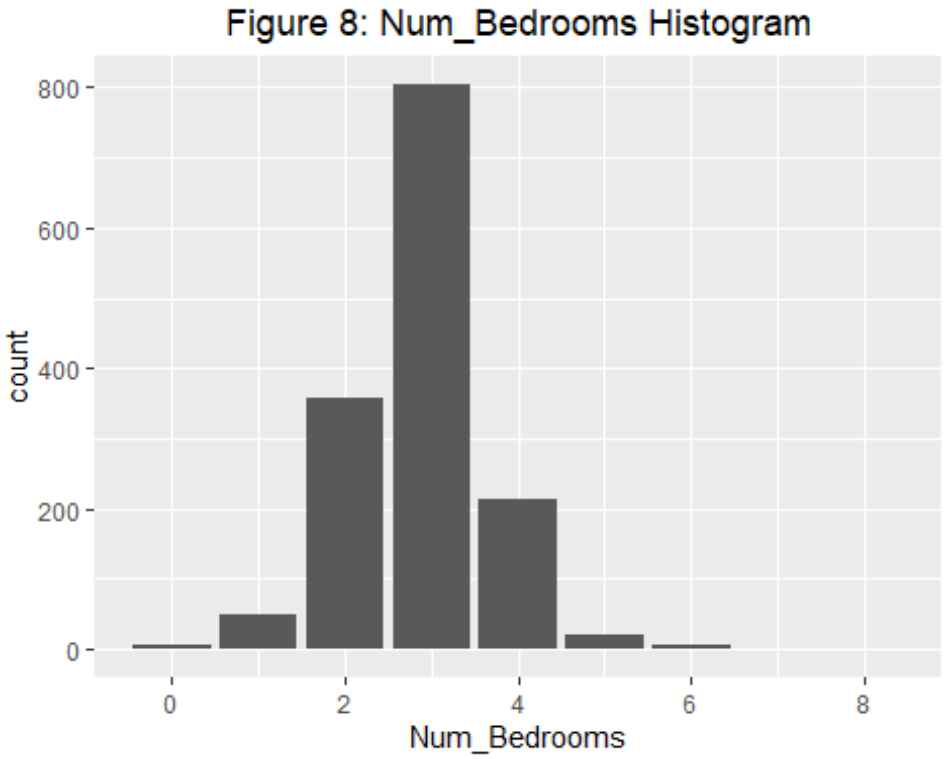
Overall Condition

Figure 7 below shows that *Cond_Ranking* is skewed to the right and that there are relatively few "poor" conditioned houses represented in the data set.



Number of Bedrooms

Figure 8 below shows that the majority of houses have two, three, of four bedrooms and that there is at least one house with eight bedrooms (this will be further investigated when I examine outliers).



Finished Basement

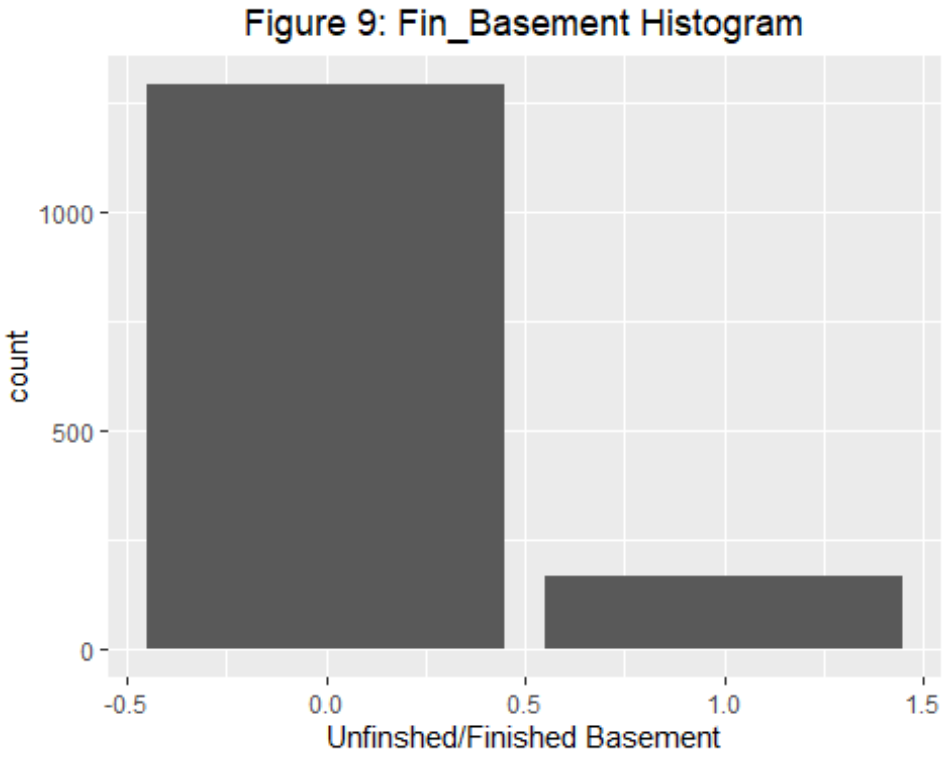


Figure 9 shows that only about 10% of the houses in the data set have finished basements.

Possible Linear Correlation

Here I examine possible linear correlation between some of my independent variables. Specifically, I am concerned that there may be a high level of linear correlation between the *Square_Footage* and *Num_Bedrooms* in a house because larger houses tend to have more bedrooms. I am also concerned with linear correlation between *Square_Footage* and *Lot_size* because large houses tend to be built on larger plots of land. Tables 1 and 2 present the covariance and correlation matrices for my independent variables.

Table 1: Var/Cov Matrix for Independent Variables

```
##                   Lot_Size Num_Bedrooms Cond_Rating Square_Footage
Fin_Basement
## Lot_Size        1.00000000   0.11968991 -0.00563627     0.26311617
0.09623564
## Num_Bedrooms    0.11968991   1.00000000  0.01298006     0.52126951
0.01135983
## Cond_Rating    -0.00563627   0.01298006  1.00000000    -0.07968587
0.08689464
## Square_Footage  0.26311617   0.52126951 -0.07968587     1.00000000  -
0.03812574
## Fin_Basement    0.09623564   0.01135983  0.08689464    -0.03812574
1.00000000
```

Table 2: Correlation Matrix for Independent Variables

```
##                   Lot_Size Num_Bedrooms Cond_Rating Square_Footage
Fin_Basement
## Lot_Size        1.00000000   0.11968991 -0.00563627     0.26311617
0.09623564
## Num_Bedrooms    0.11968991   1.00000000  0.01298006     0.52126951
0.01135983
## Cond_Rating    -0.00563627   0.01298006  1.00000000    -0.07968587
0.08689464
## Square_Footage  0.26311617   0.52126951 -0.07968587     1.00000000  -
0.03812574
## Fin_Basement    0.09623564   0.01135983  0.08689464    -0.03812574
1.00000000
```

Table 2 does show some linear correlation between square footage and the number of bedrooms, but it is only moderate (0.5213). More surprising, and promising, is the small degree of linear correlation between the other variables. Figure 8 presents a scatter plot of *Square_Footage* versus *Num_Bedrooms* to help visualize the degree of linear correlation.

Figure 10: Square_Footage/Num_Bedrooms Scatter Plot

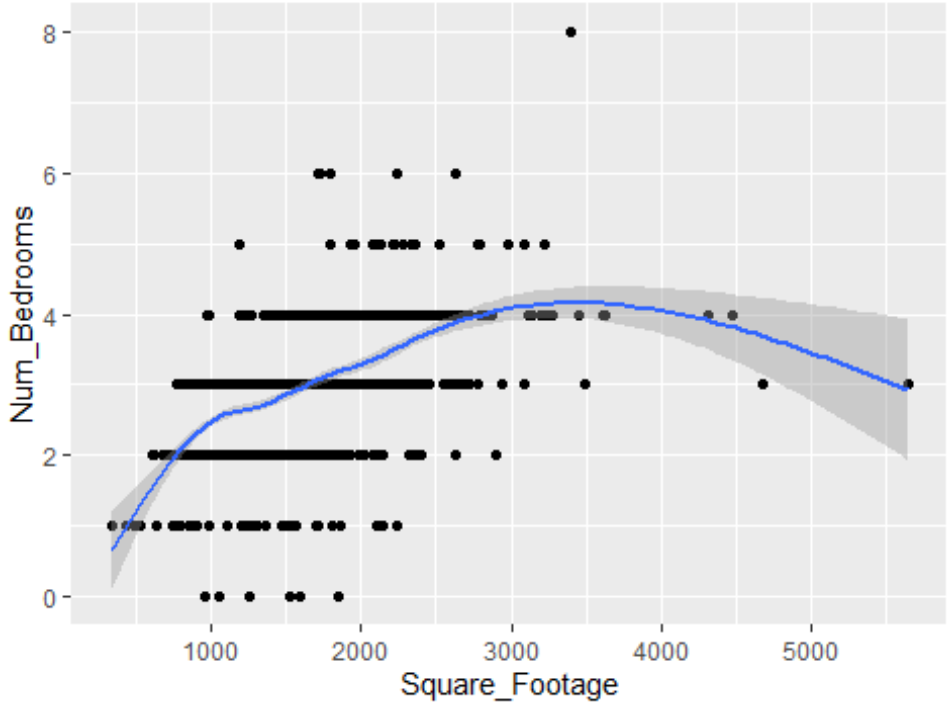


Figure 10 shows that there may be nonlinear correlation across these data. Still, it appears that there may be fairly strong linear correlation for smaller values of square footage, suggesting that removal of the outliers based on large house sizes (identified above), may allow this linear correlation to be uncovered. Later I will turn to the VIF measure to examine multicollinearity in my regression models.

Although Table 2 only shows small linear correlation between *Square_Footage* and *Lot_Size* (0.2631). I investigated my intuition about this correlation using the scatter plot in Figure 11 below.

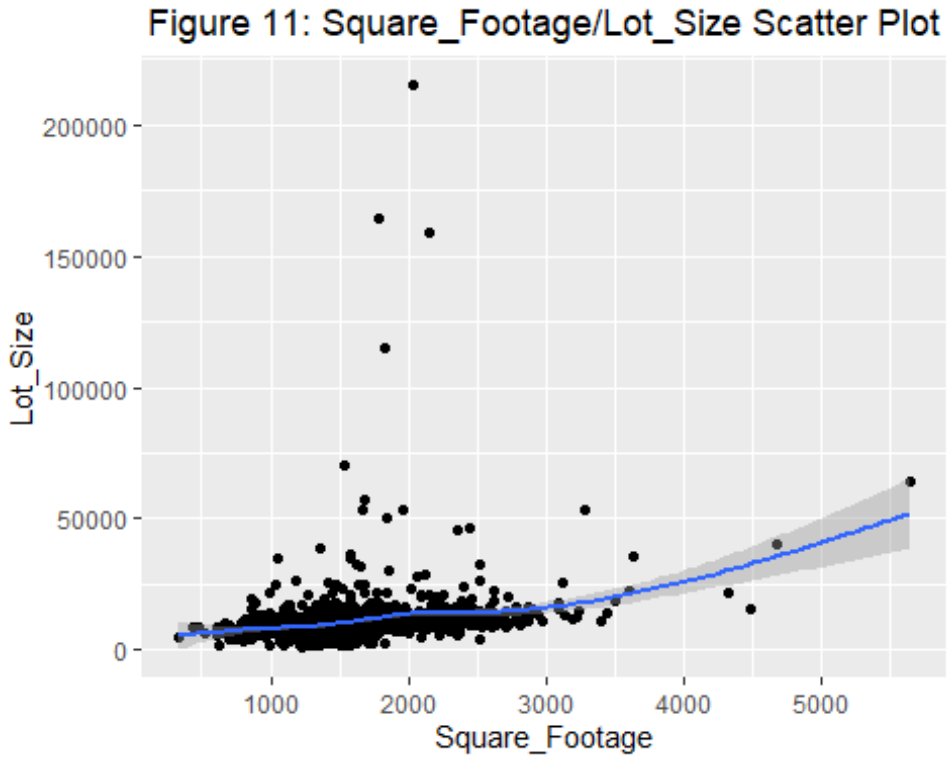


Figure 9 shows that there is small linear correlation between square footage and lot area for most house but that larger houses introduce a non-linearity to the smoothing line. this, once again, suggests the impact of outliers based on large house sizes and lot areas.

Hypotheses

The main hypotheses concern how a house's selling price is affected by a change in an independent variable (holding other things equal). My five hypotheses are:

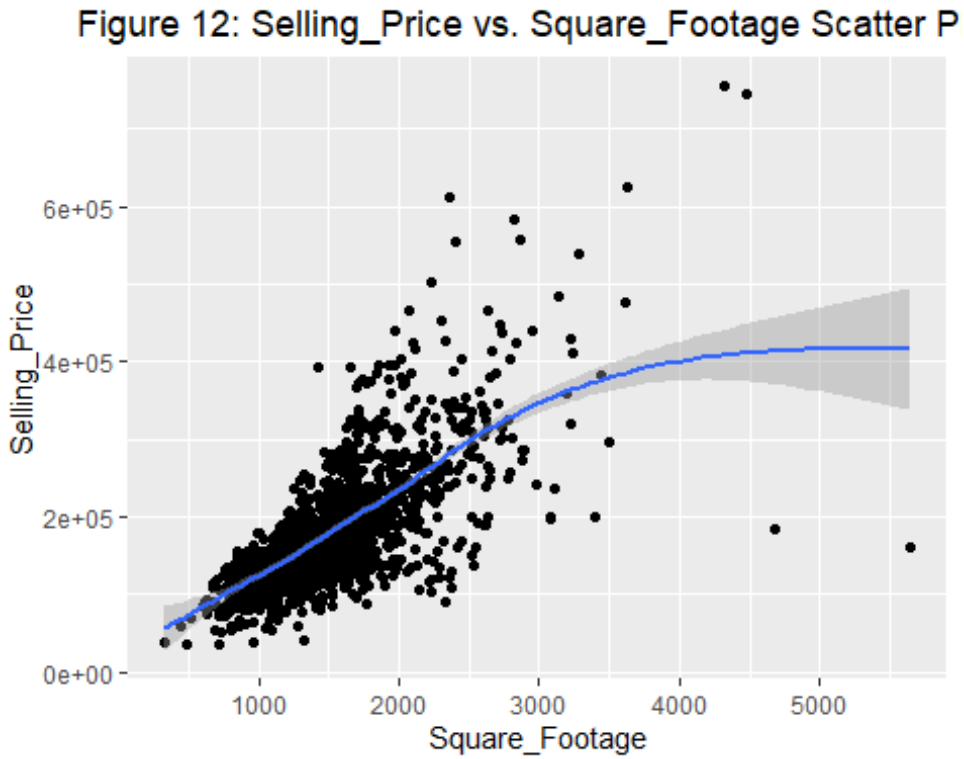
1. A larger house, in terms of square footage of living space, sells for more than a house with less square footage of living space.
2. A house built on a larger lot sells for more than a house built on a smaller lot.
3. A house with more bedrooms sells for more than a house with fewer bedrooms.
4. A house with a higher condition rating sells for more than a house with a lower condition rating.
5. A house with a finished basement sells for more than a house without a finished basement.

All of these hypotheses are based on the idea that an increase in the independent variable (*Square_Footage*, *Lot_Size*, *Cond_Rating*, *Num_Bedrooms*, and *Fin_Basement*) add value to the house and thus should result in it selling for a higher price. In the next section, I present the results from a linear regression that allows me to test these hypotheses.

Model Results

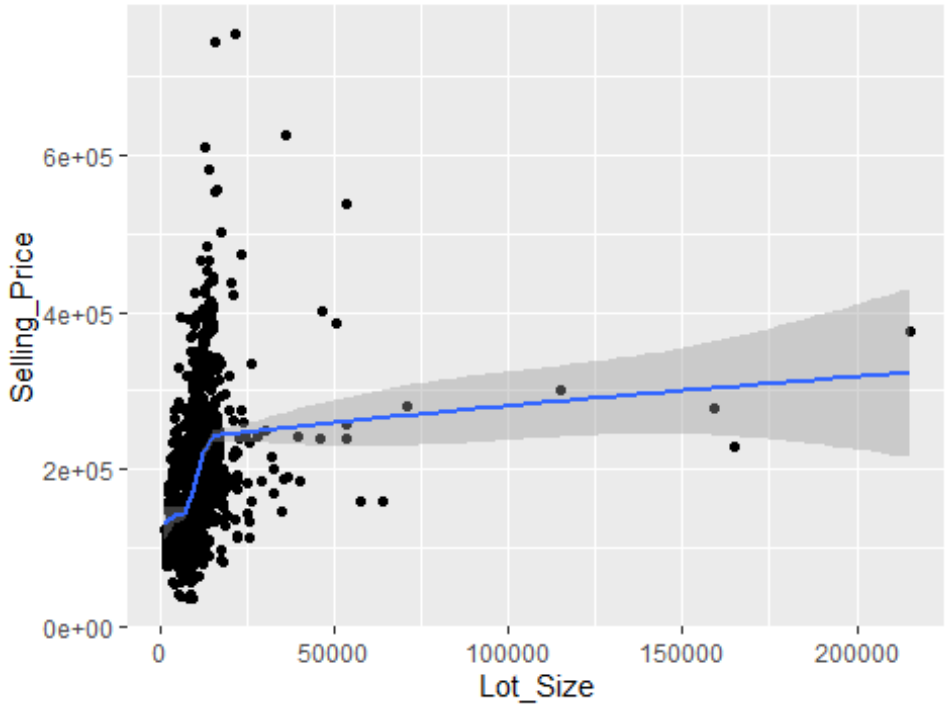
Predictive Visualizations

This section provides scatter plots that show the individual relations between *Selling_Price* (the dependent variable) and the various independent variables.



This visualization shows the strong influence of a few data points where the *Square_Footage* is far larger than most other observations. These larger houses ar "bending" the smoothing curve and introducing a non-linearity in to the relationship that is not there for most houses in the data. Once again, I will need to further investigate the presence of these observations as possible outliers.

Figure 13: Seliing_Price vs. Lot_Size Scatter Plot



Like *Square_Footage*, Figure 13 shows that several observations have very high values of *Lot_Size*, thus bending the smoothing curve and suggesting the existence of outliers. Still, there appears to be a positive relation between the two variables.

Figure 14: Seliing_Price vs. Cond_Rating Scatter Plot

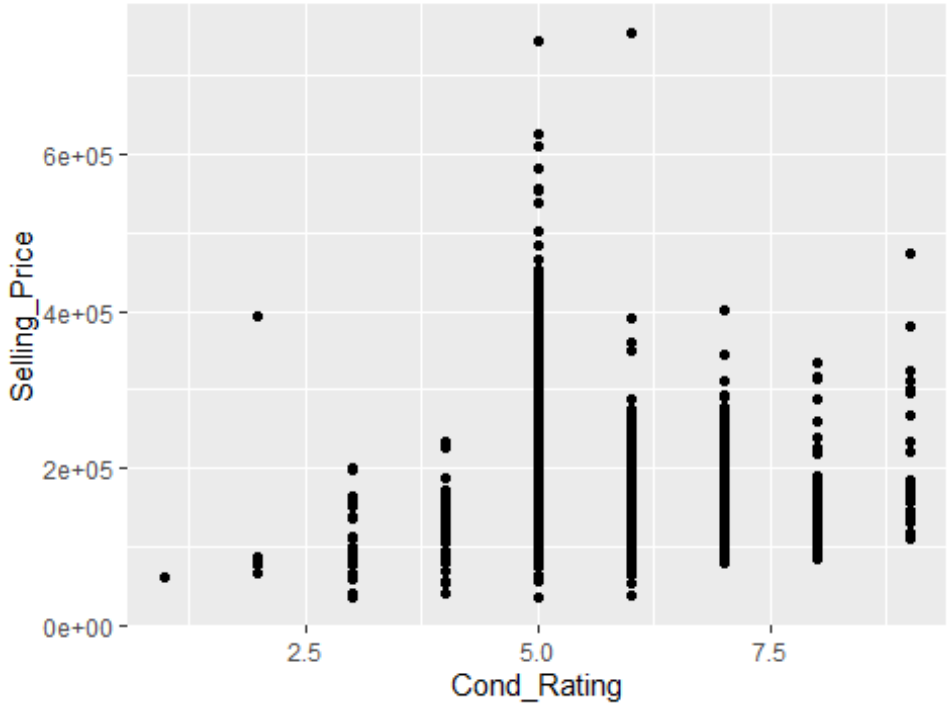


Figure 14 shows that there may be a modest positive relation between *Selling_Price* and *Cond_Rating* but it is minimal and difficult to tell if it is really positive.

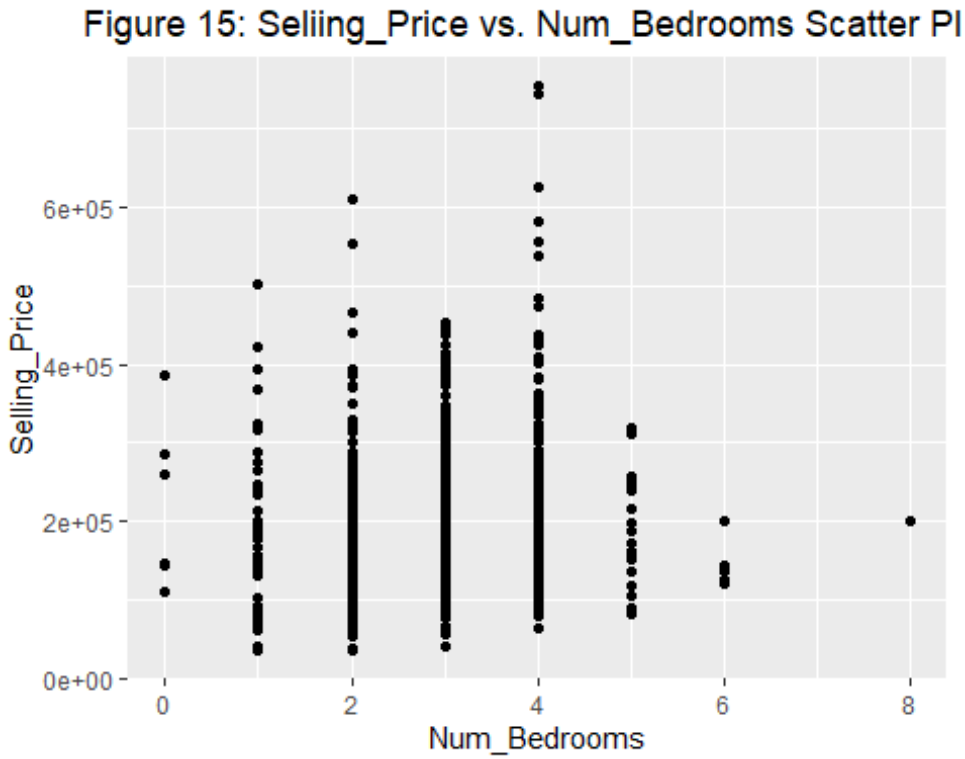


Figure 15 does not really highlight a relationship between *Selling_Price* and *Fin_Basement*. I will have to let the regression flesh this out.

Regression Model Results

This section presents results for the linear regression model

$$SellingPrice = \beta_0 + \beta_1(LotSize) + \beta_2(NumBedrooms) + \beta_3(CondRating) + \beta_4(SquareFootage) + /beta_5 * (FinBasement) + \epsilon$$

First, the model is estimated using all 1460 observations. The resulting regression equation is:

$$SellingPrice = 63540 + 0.6451(LotSize) - 26566(NumBedrooms) - 40.33(CondRating) + 125.2(SquareFootage) - 6389(FinBasement)$$

Further results can be found in Table 3.

Table 3: Initial Regression Results (outliers included)

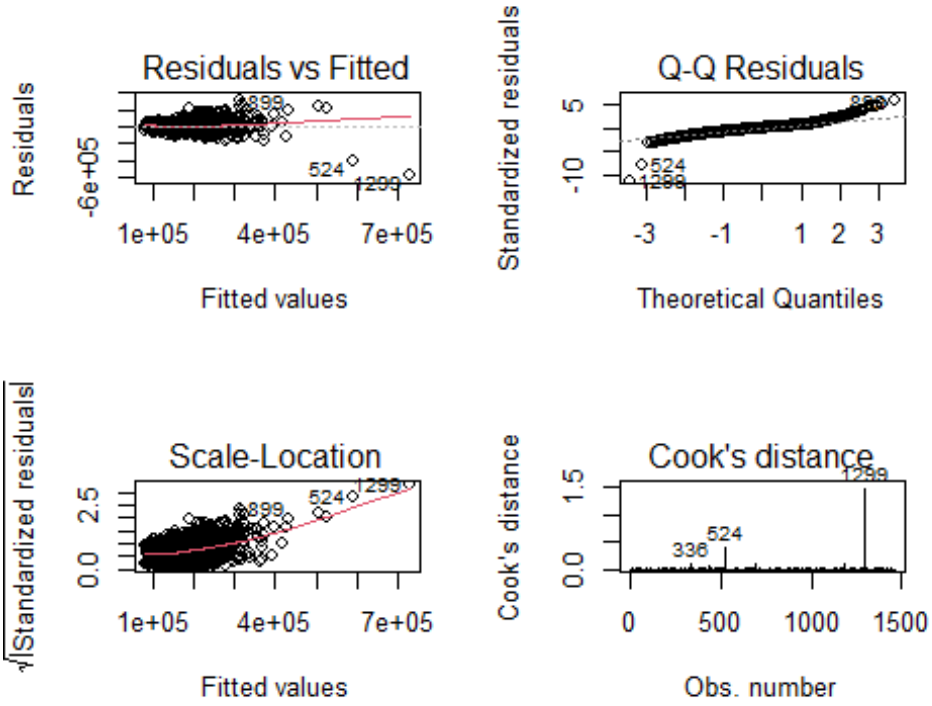
```
##
## Call:
```

```
## lm(formula = Selling_Price ~ Lot_Size + Num_Bedrooms + Cond_Rating +
##     Square_Footage + Fin_Basement, data = housing_prices3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -569349  -26834    -255   22569  298984
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.354e+04  8.820e+03   7.204 9.37e-13 ***
## Lot_Size        6.451e-01  1.438e-01   4.487 7.80e-06 ***
## Num_Bedrooms   -2.656e+04  1.982e+03 -13.405  < 2e-16 ***
## Cond_Rating    -4.033e+02  1.247e+03  -0.323    0.746
## Square_Footage  1.252e+02  3.182e+00  39.349  < 2e-16 ***
## Fin_Basement   -6.389e+03  4.368e+03  -1.463    0.144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52540 on 1454 degrees of freedom
## Multiple R-squared:  0.5641, Adjusted R-squared:  0.5626
## F-statistic: 376.3 on 5 and 1454 DF,  p-value: < 2.2e-16
```

The regression results us that the independent variables do fairly well at explaining *Selling_Price*. To begin, the adjusted R-Square value is -.5641, indicating that more than 50% of the variation in *selling_Price* is explained by the independent variables. Further, the p-value is 2.2e-16 which is very close to zero, indicating that the regression model is statistically significant. More specifically, *Square_Footage*, *Lot_Size*, and *Num_Bedrooms* are all statistically significant at a level above 99%. In addition, *Fin_Basement* is nearly significant with a p-value of 0.144. Of the five independent variables, only *Cond_Rating* is highly insignificant. I now turn to four robustness tests of the model.

The coefficients of *Square_Footage* and *Lot_Size* are, as hypothesized, positive. However, the coefficients *Num_Bedrooms*, *Cond_Rating*, and *Fin_Basement* are all negative, the opposite of what was anticipated. I suspect that these contrary results may be due to the influence of outliers. I will address this issue after examining the robustness tests from the model that includes the outliers.

Figure 16: Robustness Tests on Initial Regression (outliers included)



The four graphs in Figure 16 are insightful. First, the plot of Residuals vs Fitted Values shows some non-linearity and divergence from the horizontal line at zero. This is particularly true for the higher level of fitted values which most likely comes from the outliers discussed above. It also shows the inequality of variance as the fitted values increase, thus confirming our suspicions about heteroscedasticity from above. The Quantile-Quantile Residuals Plot diverges from the 45-degree line for low and high quantiles. This is evidence that the residuals may not be normally distributed, once again possibly caused by the outliers. The Scale-Location Plot shows the residuals are not spread equally along the ranges of predictors and is the red line is far from horizontal, indicating that there is heteroscedasticity. Finally, the Cook's Distance Plot shows that there are at least a few extreme outliers. In fact, there are more than just the three isolated by the plot. It is just that observation 1299 is so extreme, that it dwarfs other, large, outliers.

In light of the above robustness tests, I next removed the most extreme outliers and reran the regression model (I removed 89 outliers, primarily based on an abnormally large *Square_Footage* and/or *Lot_Size*). But before doing so, I present Figure 17 to show that the scatter plot of *Selling_Price* versus *Square_Footage* removes the non-linearity seen in Figure 12 as evidence that the outlier removal was effective.

Figure 17: Seliing_Price vs. Square_Footage Scatter P

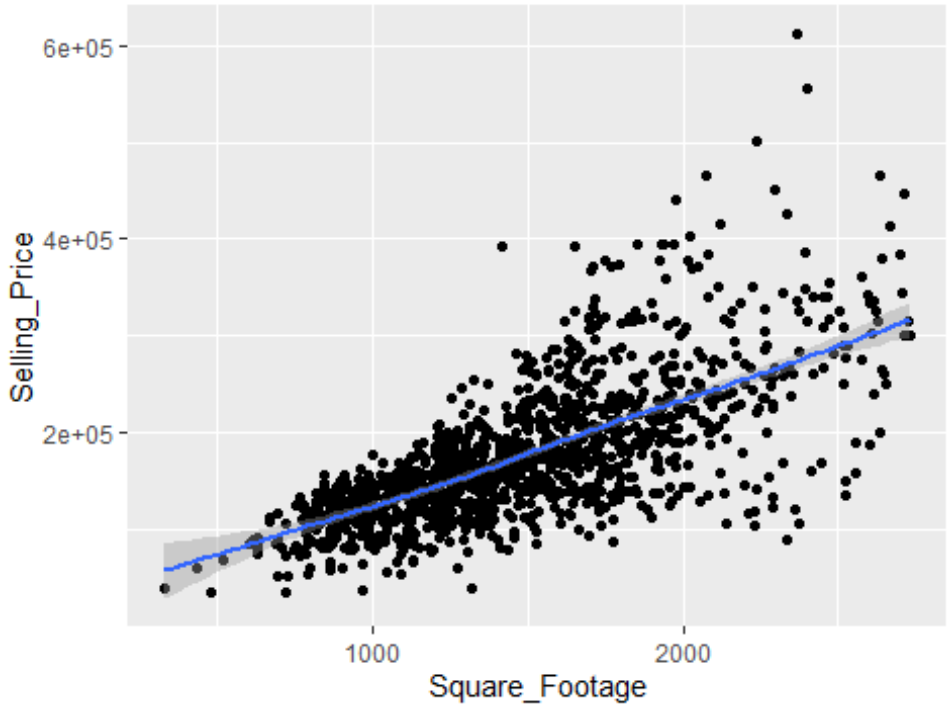


Table 4: Follow-Up Regression Results (outliers removed)

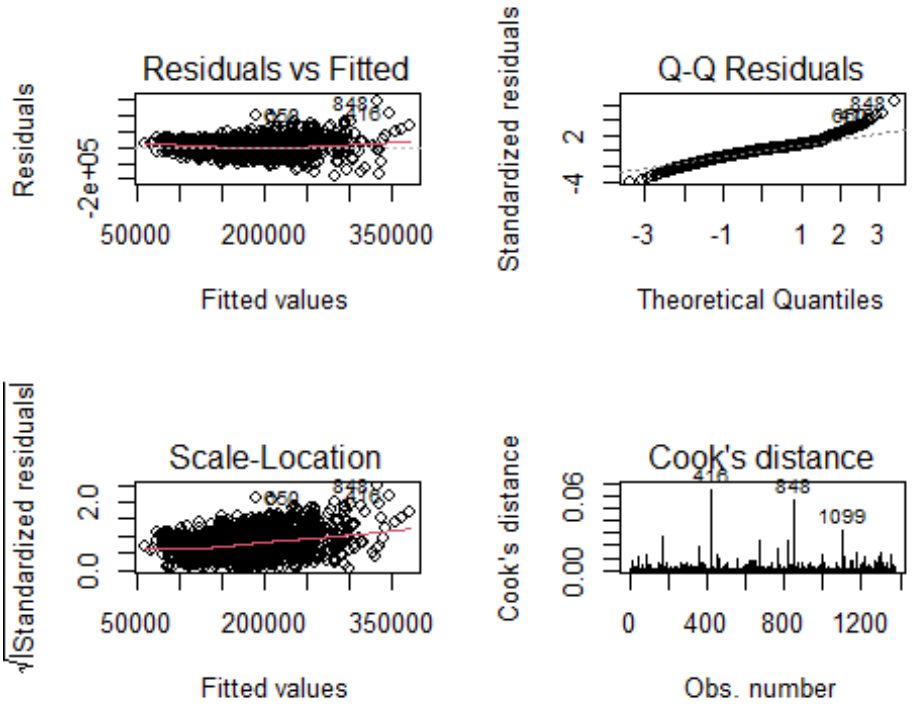
```
##
## Call:
## lm(formula = Selling_Price ~ Lot_Size + Num_Bedrooms + Cond_Rating +
##     Square_Footage + Fin_Basement, data = prices_no_outliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -176773  -23856    2034   22803  277570
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.001e+04  8.261e+03   4.843 1.42e-06 ***
## Lot_Size        4.707e+00  4.127e-01  11.406  < 2e-16 ***
## Num_Bedrooms   -3.232e+04  1.816e+03 -17.799  < 2e-16 ***
## Cond_Rating    -3.196e+02  1.107e+03  -0.289    0.773
## Square_Footage  1.267e+02  3.349e+00  37.831  < 2e-16 ***
## Fin_Basement   -2.981e+03  3.873e+03  -0.770    0.442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44410 on 1365 degrees of freedom
## Multiple R-squared:  0.6032, Adjusted R-squared:  0.6018
## F-statistic: 415.1 on 5 and 1365 DF,  p-value: < 2.2e-16
```

The new estimated regression equation is:

$SellingPrice = 40010 + 4.707(LotSize) - 32320(NumBedrooms) - 319.6(CondRating) + 126.7(SquareFootage) - 2981(FinBasement)

Table 4 shows that a 4% improvement in the adjusted R-Square and the p-value was nearly the same as with the outliers. This suggests that removing the outliers helped, but not much. In terms of the regression coefficients, although the exact estimates differ, *Square_Footage*, *Lot_Size*, and *Num_Bedrooms* are still all statistically significant at a high confidence level (> 99%) and that *Cond_Ranking* and *Fin_Basement* continue to be insignificant. In fact, the *Fin_Basement* became even more insignificant than when the outliers were included. The signs of the coefficient are also the same than before the outliers were removed.

Figure 18: Robustness Tests on Follow-Up Regression (outliers removed)



Once again, the four graphs are insightful. This time, the plot of Residuals vs Fitted Values in Figure 18 now has much less spread in the variation although there is some for higher values of the fitter values. The Q-Q Residuals Plot still deviates from the 45-degree line in the higher and lower quantiles. The Scale-Location Plot has a much more horizontal red line although there is a slight upward slope. The Cook's Distance Plot shows that there are at least a few (relatively)extreme outliers. All of this suggests that removing the outliers reduced the heteroscedasticity but that there are still issues.

Although the correlation matrices in Table 2 provided little evidence of linear correlation, I conclude my analysis of multicollinearity by examining the Variance Inflation Factor (VIF) associated with each regression.

VIF for regression that includes the outliers:

```
##       Lot_Size   Num_Bedrooms    Cond_Rating Square_Footage   Fin_Basement
##       1.088372       1.381051       1.017462       1.477116       1.021981
```

VIF for regression that excludes the outliers:

```
##       Lot_Size   Num_Bedrooms    Cond_Rating Square_Footage   Fin_Basement
##       1.198831       1.402010       1.030121       1.536920       1.021873
```

All VIF values are near 1.0, indicating that multicollinearity is not an issue in the regressions.

Conclusions

Discussion

The results of both regressions (including and excluding outliers) confirm my hypotheses that increases in *Lot_Size* and *Square_Footage* increase the *Selling_Price* of a house. My similar hypotheses regarding *Cond_Rating* and *FinBasement* were not supported because neither's coefficient was not statistically significant. Further, both were actually negative rather than positive as hypothesized. Finally, not only was *Num_Bedrooms* statistically significant, it was surprisingly negative, a complete rejection of my fourth hypothesis. In essence, this means that increases in the size of the livable area and houses on larger lots will sell for more. Practically, this suggests that there is value in putting an addition on your house or perhaps obtaining contiguous property. The estimate for square footage is particularly enticing as it suggests that a 100 square foot addition to a house will add an estimated $12,670.

Business Value (Prescriptive Recommendations)

The above example of a 100 square foot addition to a house provides an excellent example of how a business could utilize my results in a decision making capacity. That is, if the cost of an addition to a house is less than $12,670, it suggests that the addition could be a good idea, but that for higher costs, it should be prohibitive. A similar judgement could be made about whether or not to finish an unfinished basement although under the current regressions, I cannot make a definitive claim that finishing a basement would add any value to a house since that coefficient was insignificant. But the applications to business do not stop there. As mentioned at the beginning of the report, a Realtor can input the characteristics of a house into my regression equation in order to forecast the price at which the house will sell. Theoretically, it will also allow agents to help sellers make decisions about updates/upgrades to the house before putting it on the market. For instance, if the cost of improving the condition of the house is less than the estimated benefit of a higher condition rating, this would signal that improving the condition is cost effective. Unfortunately, from a practical standpoint, the regression coefficient on *Con_Rating* was insignificant so any decisions based on that coefficient would be misleading.

In summation, The key takeaways from this report ar that the size of a house and the size of the plot of land it is built on very likely influence the value of the house in a positive direction. Other influences are questionable due to the insignificance of some coefficient. This engenders further research on the issue, particularly that tat adds independent variables and deals with heteroscedasticity. My final recommendation is to apply the regression results with caution and to further investigate what causes changes in housing prices.

Limitations and Future Research

The main limitation of the first model is the presence of outliers involving larger houses and very large plots of land. This was evidenced by the data visualization and verified by the Cook's Distance Test. The main limitation of the second regression is that it only applicable to real estate in Iowa that are less than 2747.5 square feet and/or sit on a lot less than 17,674 square feet (these values were selected because they are 1.5 times Q3 for each). Limiting in both regressions is that only five independent variables were used and the coefficients on three of them were statistically significant. Still, I do not think anything restricts my inferences about causality in my regressions. Each of the five independent variables were chosen exactly because of the expectation that they directly influence the selling price of a house.

Future research could remove the insignificant variables and add more variables that should influence the value/selling price of the house. These data are not limited to the initial data set either. One might also collect data on crime rates and school quality in a house's neighborhood. Those are significant factors that affect demand for a house and would likely improve the adjusted R-Square value. Additionally, improved estimates would probably result if the finished basement variable could be expanded to include a partition into finished basements, unfinished basement, and no basement. This would require the creation of two dichotomous variables to handle all possible cases and still avoid the "dummy variable trap" and improve the overall model. Finally, any future research should definitely address the remaining outliers and heteroscedasticity.