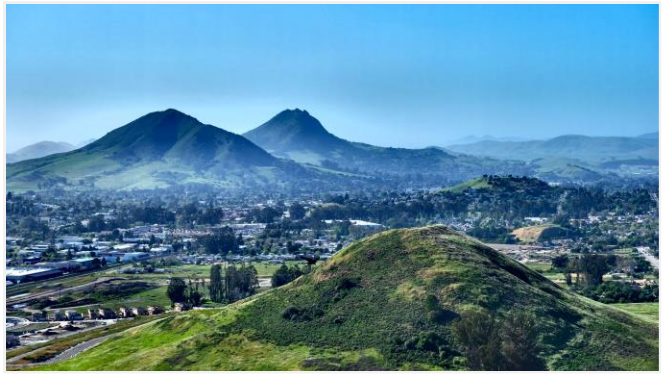


Residential Regression

What drives housing prices in SLO?

Brendan Callender, Martin Hsu, and Rachel Roggenkemper

College of Science and Math
Statistics
California Polytechnic University San Luis Obispo

Abstract

Anyone who dabbles in real estate or anyone who watched *Fixer Upper* can tell you the main attraction and driving point of a house's price is its location. Although we acknowledge that location is an imperative part of buying a home, we want to eliminate the effects of location to try and find what other aspects affect the price of a home. In order to do this, we randomly selected data off of Zillow ourselves, choosing from data from properties sold in San Luis Obispo (SLO), California in a 90-day time span from February 2022 through April 2022. We collected variables such as sold price, square footage, number of beds, baths, and parking spaces, the year the house was built, lot size, home type, and whether or not they had a pool, cooling, or heating on each property in our data set. Ultimately, we found that the selling price of a home can be attributed to three main factors: the square footage of a home, the square footage of the lot size, and the type of home. We found that these three variables are significant predictors when determining how much a home sells for.

Contents

II. Introduction

As the price of homes increases, the thought of becoming a future homeowner becomes scarier and scarier. House prices vary due to an unthinkable number of variables that seem nearly impossible to keep track of. This begs the question: What drives the price of houses and how do you know you are getting a good deal? If you asked a real estate agent, they would probably reply that the price of a home is driven by, “location, location, location.” In fact, studies such as the one conducted by the Royal Institute of Technology[1] have investigated which aspects of location affect real estate prices using regression analysis. In order to try to eliminate the effects of location, we will be fitting our regression model to a random sample of recently sold homes located here in San Luis Obispo. The variables that we will be using to predict sale price include both quantitative and categorical variables. The quantitative variables include the size of the home in square feet, the number of bedrooms, the number of bathrooms, the year the house was built, the lot size, and the number of parking spaces. The categorical variables include whether or not the housing has cooling, heating, or a pool and also includes the type of the home. These variables fall under the umbrella of building properties. In fact, one study conducted by the Hong Kong University of Science and Technology and published in *Scientific Programming*[2] noted that building properties such as garage capacity, backyard space, and having a swimming pool all played key roles in predicting price. Looking at previous studies played a key role in deciding which variables to use to attempt to accurately predict the prices of homes.

Data Ethics

Data is extremely powerful and can be used to educate others on decision making. It is extremely important to consider those affected by the conclusions drawn from data. For example, when working with data for diseases such as COVID-19, it is extremely important to understand that every observation in the data represents a loved one who has sadly passed away. These considerations should help shape the conclusions drawn from the data as well as the presentation of the conclusions. Due to the objective of the study being to predict housing prices in San Luis Obispo, the data ethics side of our report is much less trivial than studies for heavy topics like COVID-19. However, there are still many important considerations for us. Firstly, our model may benefit or harm homeowners in SLO by inflating or deflating the value of their homes by our model not being accurate in reality. Secondly, our model may harm those looking for affordable housing and may over-value homes that should be within reach in reality. With these considerations in mind, please view our conclusions as a practice exercise in using linear regression and not as a true model for the value of homes. As famous British statistician George Box says, “All models are wrong, but some are useful.”

III. Materials and Methods

This study was an observational study of San Luis Obispo (SLO), California properties sold within a certain time period. On April 29th, 2022, we sourced our data by manually scraping the Zillow listings of sold properties in SLO. The process first involved opening a search query into SLO properties on the Zillow website, then filtering to non-empty lot properties that were sold within the past 90 days. Each selected listing was clicked on one-by-one and the information on each listing relevant to the explanatory variables was processed and converted to the appropriate units if necessary. For instance, size of the lot was measured in both square feet and acres, so observations with lot size in acres needed to be converted to square feet. Finally, these values were input to an Excel spreadsheet in the appropriate row and column based on observation and variable.

While the process was long and prone to human error, other methods of data scraping, such as using automated scripts written in Python or other scraping software, were unable to be developed in a timely manner due to the complex layout of the Zillow website.

In all, 144 listings were randomly selected. This was achieved by ordering the related listings in a manner unrelated to time, assigning each listing a number based on its order in the sequence, and randomly selecting the numbers with a random number generator.

The observational unit was each property, selected based on if it was sold in SLO within the 90 days before April 29th, 2022. The response variable was the price, in nominal $USD, that the house was sold at (SoldPrice). The explanatory variables recorded included the following:

Sqft – The interior size of the home in square feet
Bed – The number of bedrooms in the home
Bath – The number of bathrooms in the home
YearBuilt – The year the home structure was built
LotSize – The size of the property lot in square feet
Parking – The number of designated parking spaces
Cooling – Whether or not the house has air conditioning (yes, no)
Heating – Whether or not the house has a heating system (yes, no)
Pool – Whether or not the house has a swimming pool (yes, no)
HomeType - Type of property (SingleFam, Condo, Manufactured, Duplex, MultiFam, Townhouse, Triplex, MobileManufactured, PlannedDevelopment)
HomeTypeNew – Variable mutated from HomeType, with four types of property instead of 9 (SingleFam, Condo, Manufactured, Other), where Other is a combination of similar levels with the smallest sample sizes (Duplex, MultiFam, Townhouse, Triplex, MobileManufactured, PlannedDevelopment). For reasons of balance, this is the version of HomeType we used.

IV. Splitting the Data

In order to confirm the predictive ability of the model – that is, to show whether or not this model can be used to predict housing prices for SLO on any given 90-day period, we must perform a model validation. To prepare for this process, we split the sample data of SLO properties sold within the past 90 days before April 29th, 2022, and selected twenty-nine data points, or roughly 20% of the sample, to set aside for validation using a randomized selection process in R. These data points, each representing a single property, were subset into a separate test data frame and were not used in the regression. The remaining 116 data points were kept for use in the regression analysis and stored in their own training data frame. To keep the randomly selected data consistent for each time the code was run, we set the randomization seed in R to seed 678.

The data were later subset to properties with a lot size of 16,000 square feet or under, and then re-split into 113 observations in the training data and twenty-eight observations in the test data, again under seed 678. A justification and process for this data subsetting is outlined in subsequent sections.

V. Data Visualization

Before beginning the process to fit the model to the data, we wanted to analyze our dataset visually to gain intuition and insight into the relationship between variables and inspect any qualities of interest. This is an important first step to building a robust model. To this end, we generated scatterplot matrices of numerical data, interaction plots, histograms of numerical data, and 3D scatterplots of two numerical variables as predictors of the selling price of homes.

A. Scatterplot Matrix

Figure 5.1 provides a preliminary visualization of the relationship between numerical variables in the sample data of SLO properties sold within the past 90 days before April 29th, 2022, through a scatterplot matrix. This visualization is of all the data points, before inspecting and subsetting the data. It appears that the variable most strongly associated with the response variable, the price at which the property was sold (SoldPrice), is the square footage of the interior of the property structure (Sqft). This association appears to be the strongest linear association. This claim can be supported by taking a look at the correlation matrix, which shows that the correlation coefficient between sold price and square footage is 0.82, which implies a strong linear association. It additionally does not appear to violate assumptions required to run a linear regression, namely linearity and homoscedasticity. The variables most weakly linearly correlated with the response SoldPrice initially appear to be either the number of parking spaces on the property (Parking) or the year the house was constructed (YearBuilt).

The strengths of these associations are confirmed by the correlation matrix in Table 5.1. The correlation coefficient between sold price and number of parking spaces is 0.36, which implies a fairly weak linear association. Similarly, the correlation coefficient between sold price and year built is only 0.05, which implies that there is practically no linear association at all between those two variables.

Figure 5.1: Scatterplot Matrix of Numerical Variables

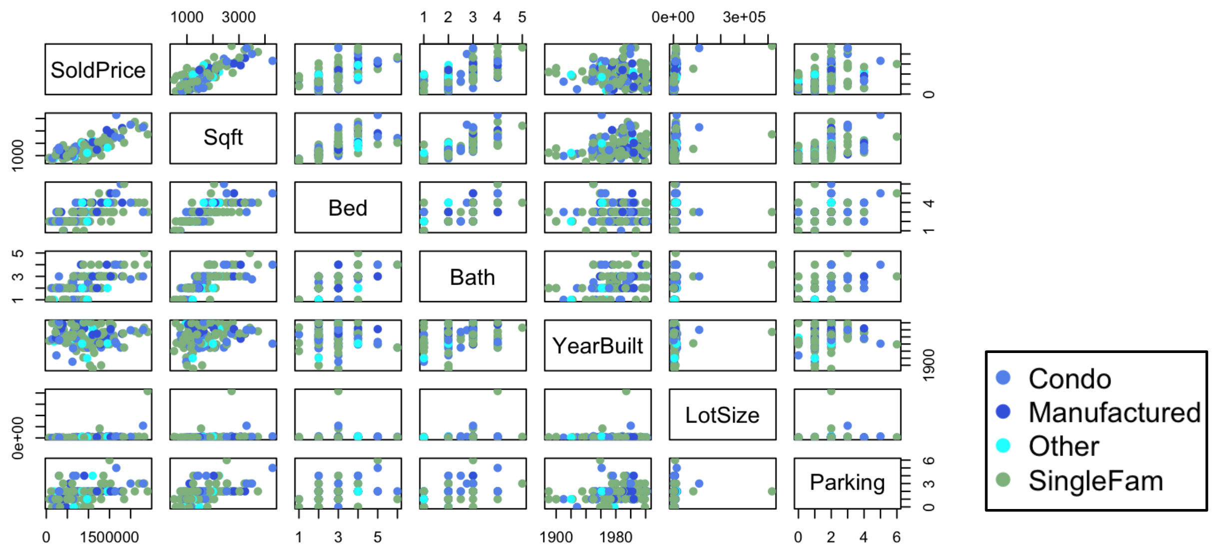


Table 5.1: Correlation Matrix of Numerical Variables

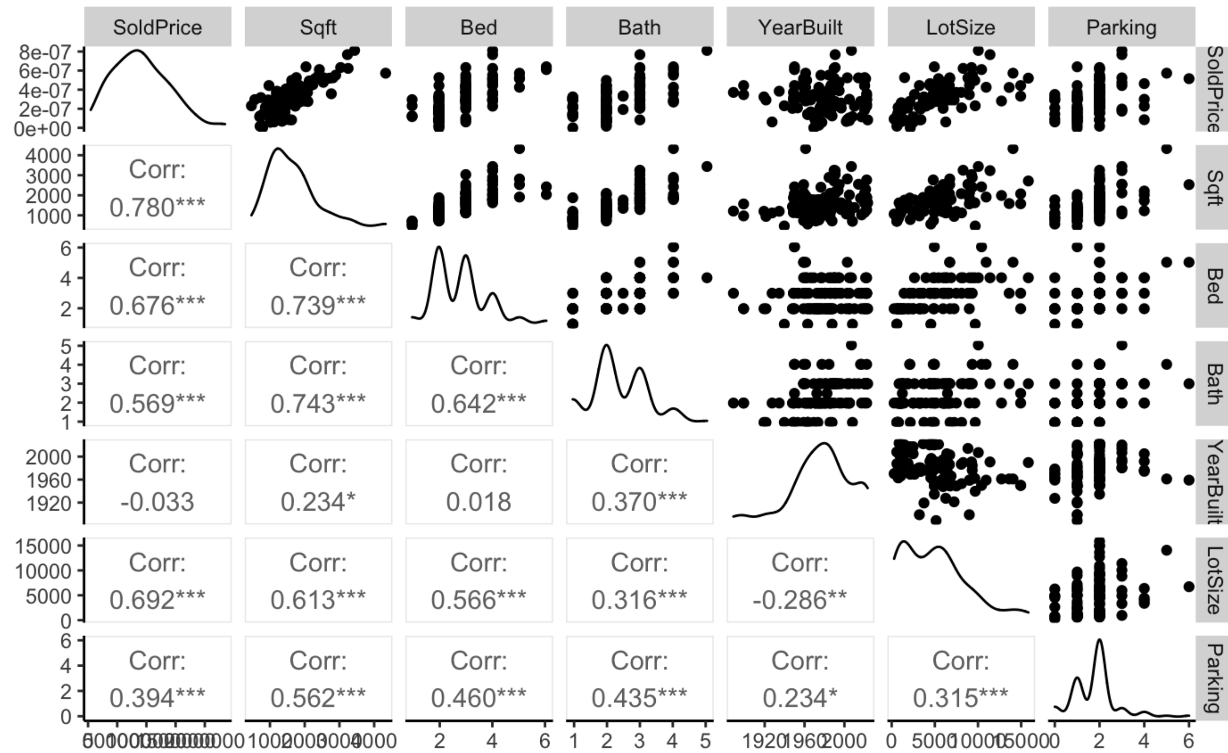
	SoldPrice	Sqft	Bed	Bath	YearBuilt	LotSize	Parking
SoldPrice	1.00	0.80	0.64	0.60	-0.01	0.37	0.36
Sqft	0.80	1.00	0.71	0.76	0.24	0.23	0.52
Bed	0.64	0.71	1.00	0.61	-0.01	0.08	0.39
Bath	0.60	0.76	0.61	1.00	0.36	0.21	0.43
YearBuilt	-0.01	0.24	-0.01	0.36	1.00	0.02	0.21
LotSize	0.37	0.23	0.08	0.21	0.02	1.00	0.07
Parking	0.36	0.52	0.39	0.43	0.21	0.07	1.00

Most, if not all explanatory variables appear to be roughly positively correlated with SoldPrice. This turns out to be correct, as nearly all of the coefficients in our correlation matrix from Figure 5.1 are positive correlation coefficients, showing that they all have varying levels of positive linear associations with one another. However, associations of SoldPrice with explanatory variables Bed, Bath, YearBuilt and Parking appear to violate the linear regression assumption of homoscedasticity. Additionally, the variable LotSize, the total area of the property in square feet, appears to have two unusual observations in the form of outlying points on plots with respect to LotSize. These appear to originate from a single-family residence with LotSize equal to 416,869 square feet, another single-family residence with LotSize equal to 108,900 square feet, and another single-family residence with LotSize equal to 84,071 square feet. As a result, the plots where LotSize is included are skewed and from Figure 5.1 it is inconclusive as to whether there exist correlations between LotSize and any of the other variables in the data.

Explanatory variables that may be positively correlated with each other include Sqft and the number of beds (Bed), Sqft and the number of bathrooms (Bath), and Sqft and the number of parking spaces (Parking). This is supported by the correlation matrix, as square footage and number of beds has a correlation coefficient of 0.72, square footage and number of bathrooms has a correlation coefficient of 0.75, and square footage and number of parking places has a correlation coefficient of 0.53. This makes sense, as it is plausible larger houses or houses with a larger footprint would tend to have more bedrooms, bathrooms and garage or parking spaces. Likewise, Bed and Bath have a relatively strong correlation, which can also be seen in the correlation matrix because bed and bath have a correlation coefficient of 0.61. Interestingly, there may be a curvilinear relationship between YearBuilt and SoldPrice as well as YearBuilt and Sqft, though creating a linear model for these relationships would require a transformation due to the violation of regression assumptions.

With the three most outlying LotSize points temporarily removed for the sake of inspection, the relationship between LotSize and the other variables becomes much more apparent in the scatterplot matrix, as seen in Figure 5.2.

Figure 5.2: Scatterplot and Correlation Matrix, with Outlying LotSize Observations Removed



In Figure 5.2, LotSize appears to be most strongly linearly correlated with SoldPrice and Sqft. These correlations are positive, which makes sense – it is plausible that homes with a larger footprint would be built on larger lots, which would lead to a stronger positive correlation between LotSize and Sqft. We would then expect LotSize to display a correlation to the response variable, SoldPrice, similar to Sqft. As an interpretation, since homes with a larger footprint tend to be built on larger lots and additionally tend to be sold at a higher price, we would expect homes with a larger lot area to also be sold at a higher price. This is supported by an inspection of both the scatterplot and correlation matrix in Figure 5.1, since the correlation coefficients between LotSize and SoldPrice is 0.69 and the correlation coefficient between LotSize and Sqft is 0.61, both similar positive correlations.

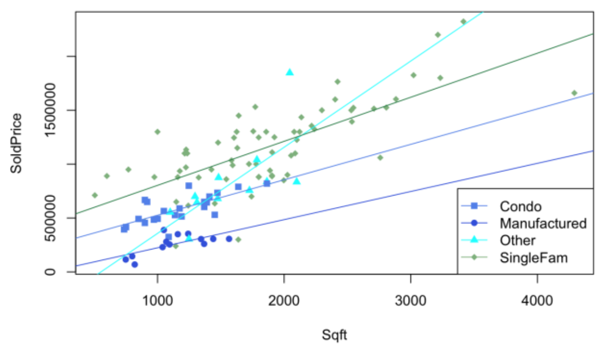
Most surprisingly, there appears to be a moderate to weak negative linear correlation between YearBuilt and LotSize, more specifically with a correlation coefficient of -0.29. Even more interestingly, though LotSize and Sqft are positively correlated, Sqft appears to have a moderate to weak positive correlation with YearBuilt with a correlation coefficient of 0.23, as opposed to negative correlation of LotSize with YearBuilt. In other words, based on the sample data, it appears that though later built SLO homes tend to be larger, they also tend to be built on smaller lots, even though larger homes tend to be built on larger lots.

As a result, we decided that these two numerical predictors, LotSize and Sqft, should be investigated further. Furthermore, we decided that it may be important to handle observations outlying in LotSize in some way.

B. Interaction Plot

To inspect whether the effect of square footage (Sqft) on the price the home was sold at (SoldPrice) changes depending on the type of home (HomeTypeNew), we generated an interaction plot seen in Figure 5.3.

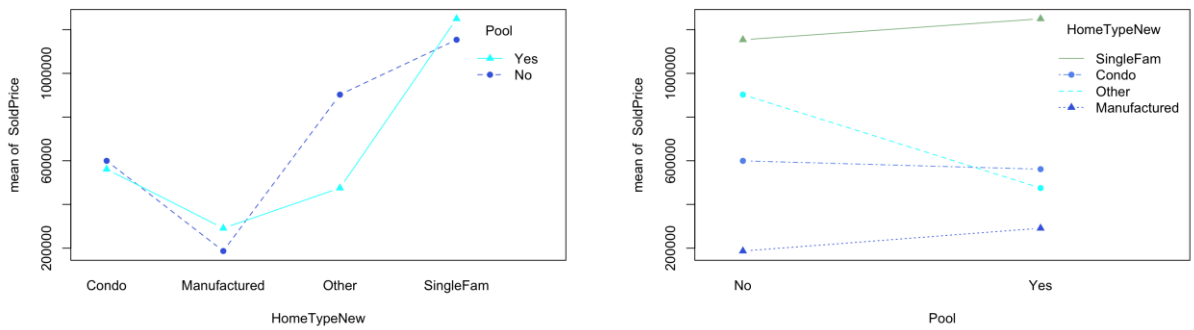
Figure 5.3: Interaction Plot of Sqft as a Predictor of SoldPrice, by HomeTypeNew



Generally, as square footage increases, the sold price of a home increases for all types of homes. The rate of increase of the Condo, Manufactured, and SingleFam levels of HomeTypeNew do not appear to differ, evidenced by the nearly parallel lines. Therefore, for these three levels, HomeTypeNew does not have an effect on Sqft. However, the least squares line for the level Other intersects the least squares lines of the Condo, Manufactured and SingleFam levels, implying that whether or not the home falls under the "Other" category has an effect on Sqft. Therefore, we concluded that an interaction between square footage and type of home may be a parameter of interest to include in our linear model. In other words, as the square footage of the home increases, the selling price of the home always increases, but this rate of increase may depend on the type of home.

Another interaction of interest between categorical variables that we investigated was between whether or not a home had a pool (Pool) and type of home (HomeTypeNew).

Figure 5.4: Interaction Plot of Pool and HomeTypeNew



As seen above in Figure 5.4, the behavior of type of home changes depending on whether or not the house has a pool. While not evident in the plot on the left where HomeTypeNew is in the x-axis, it is clear from the plot on the right that the line for the "Other" home type is not parallel to the others, resulting in the conclusion that an interaction may be present. As a result, an interaction between Pool and HomeTypeNew may be a predictor of interest to include in the model, if found significant.

In general, "Other" homes appear to have differing behavior from the other types, which may be due to the aggregate nature of the category.

C. Histograms of Quantitative Variables

To find significant descriptive properties of our quantitative variables, we created and analyzed distributions of two numeric variables of interest, the square footage of the structure (Sqft) and the square footage of the whole property lot (LotSize).

From Figure 5.5, we can see that the distribution of square footage is approximately normal but is also right skewed. This shows that while most housing is around 1,000-2,000 square feet, there are a few outliers on the larger side, implying that the mean square footage of a home structure is larger than the median.

When we see the distribution split by type of home (HomeTypeNew) in Figure 5.6, we can see a rough explanation of why the values of Sqft are skewed right. Each home category reveals a relatively symmetrical normal distribution, but an overwhelming majority of homes in the data are single family, and the other categories (Condo, Manufactured, Other) are centered around lower property sizes than the single-family homes. This imbalance should be important to keep in mind when interpreting the model.

Figure 5.5: Distribution of Sqft

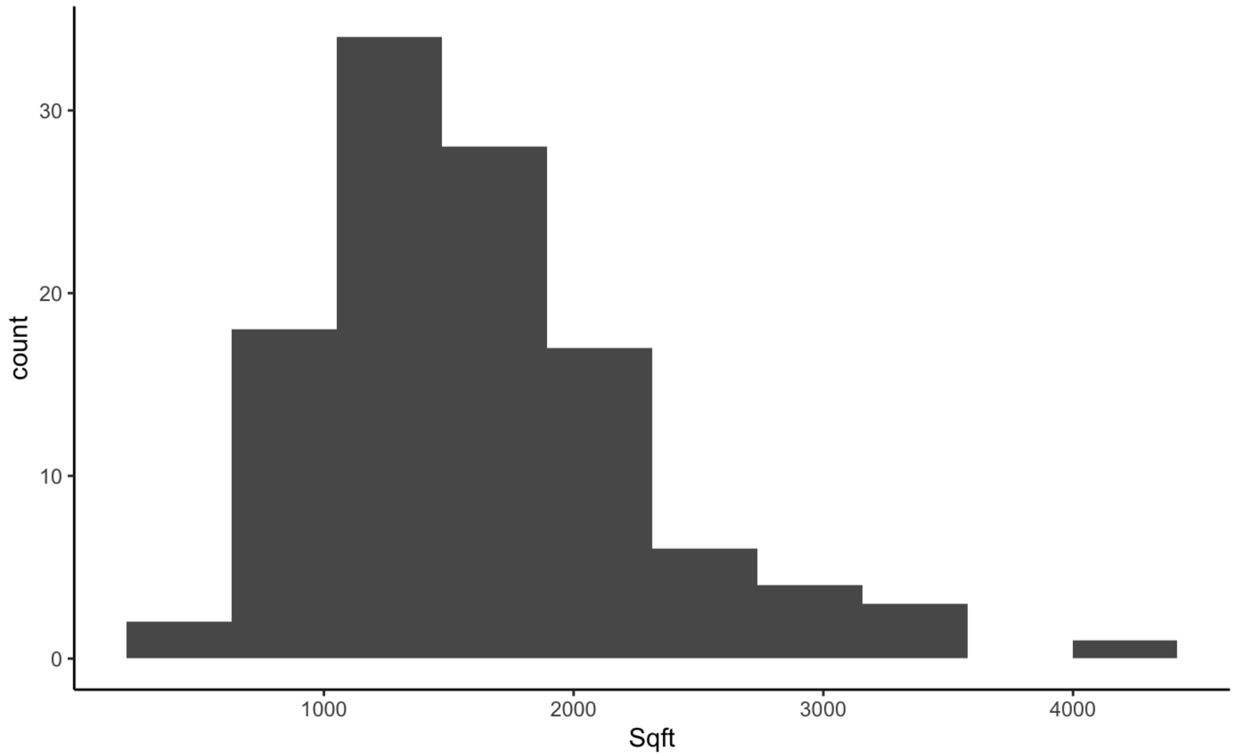


Figure 5.6: Distribution of Sqft, by HomeTypeNew

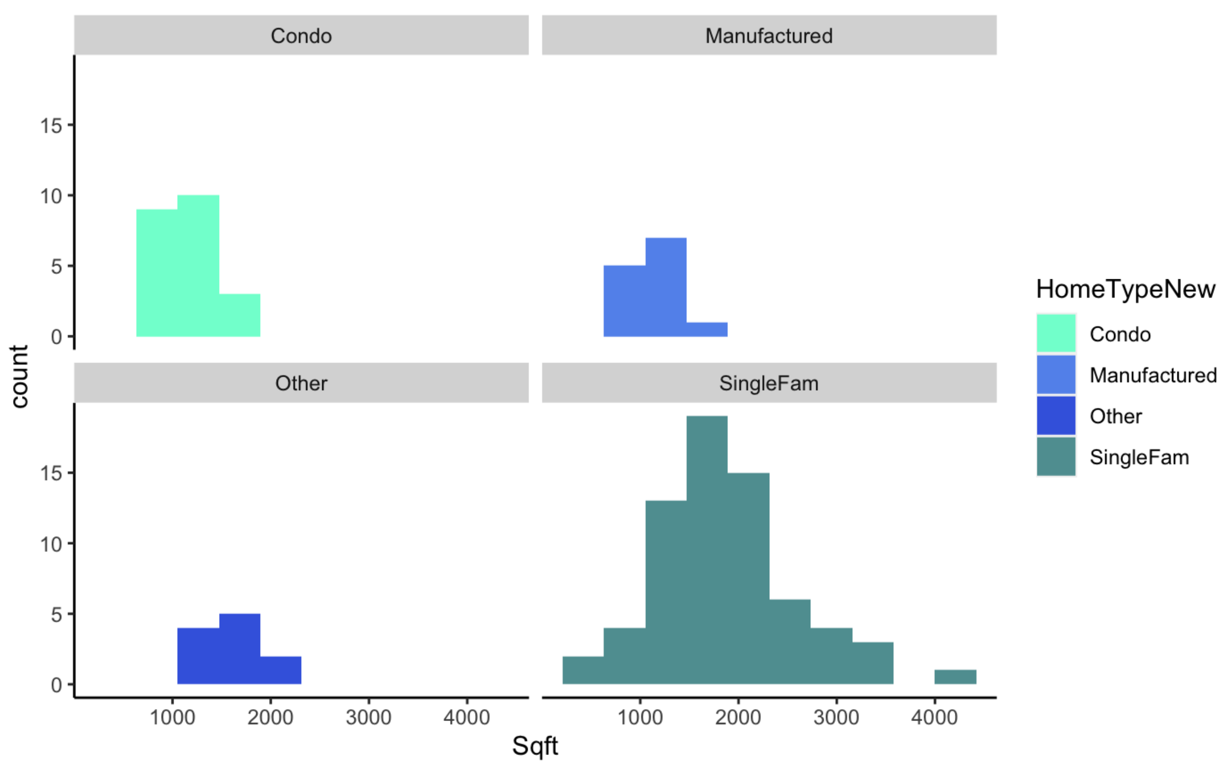


Figure 5.7: Distribution of LotSize

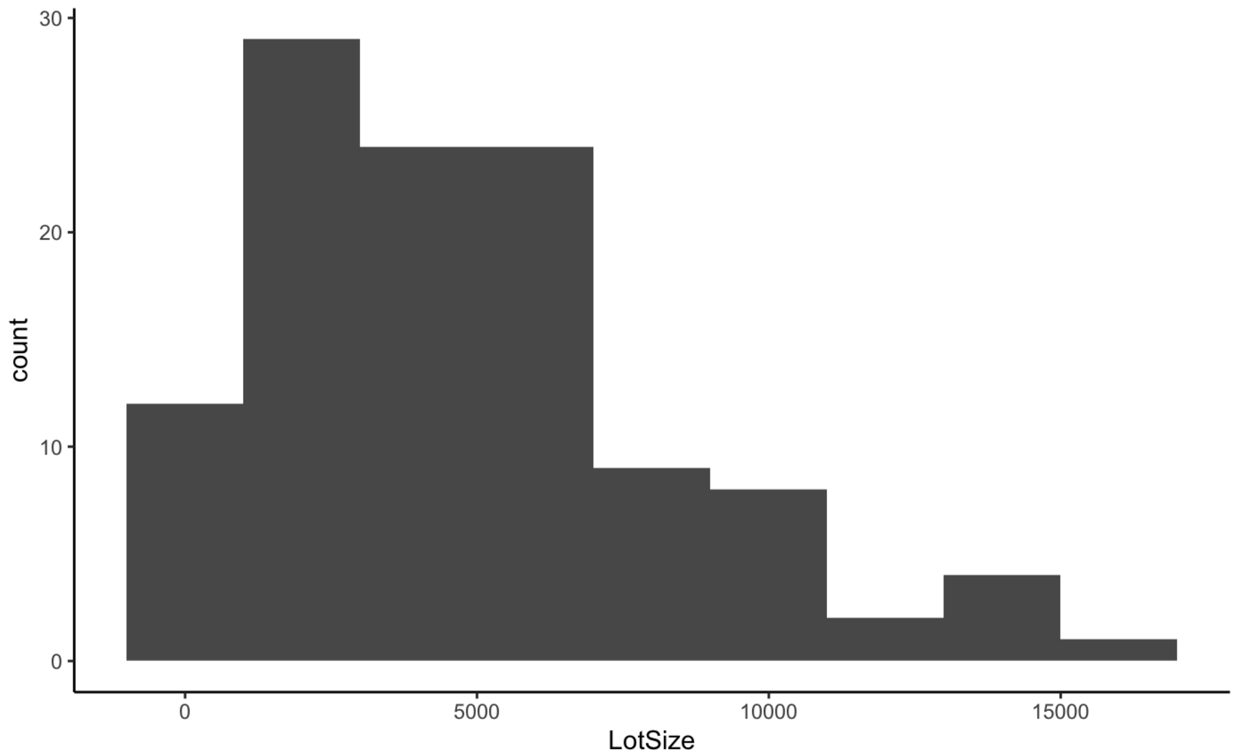
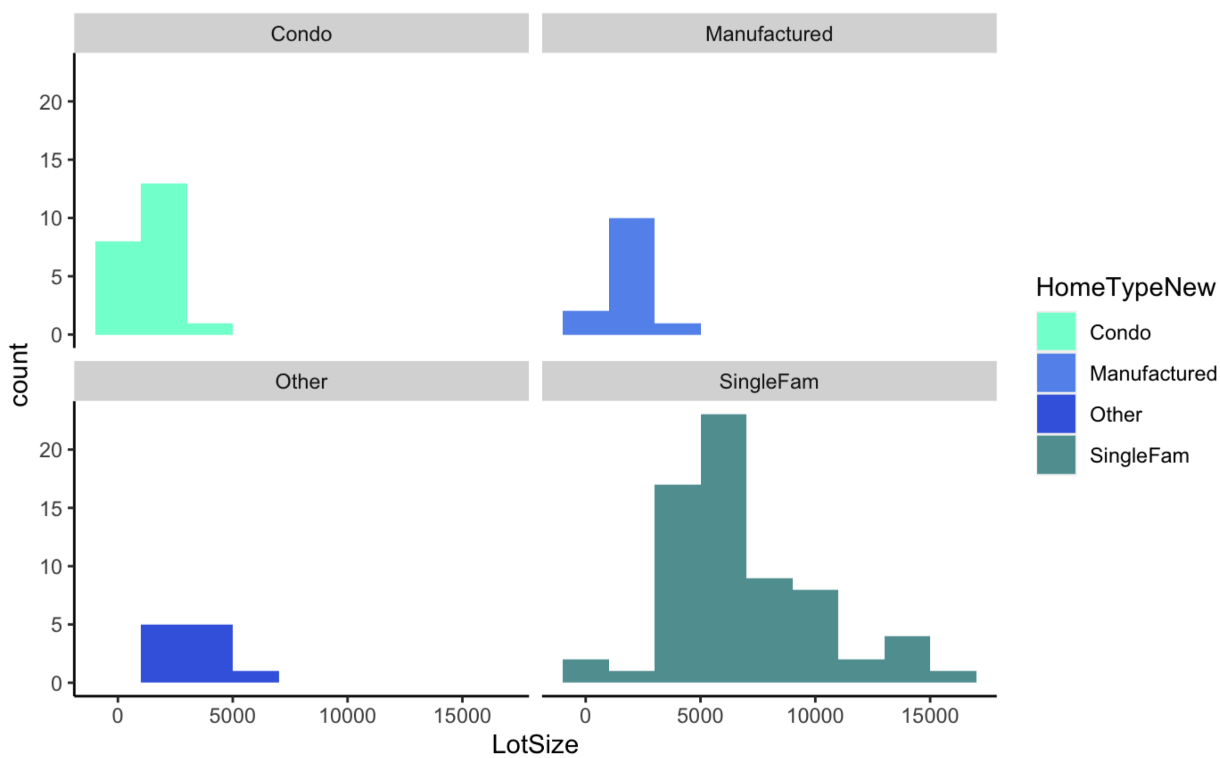


Figure 5.8: Distribution of LotSize, by HomeTypeNew

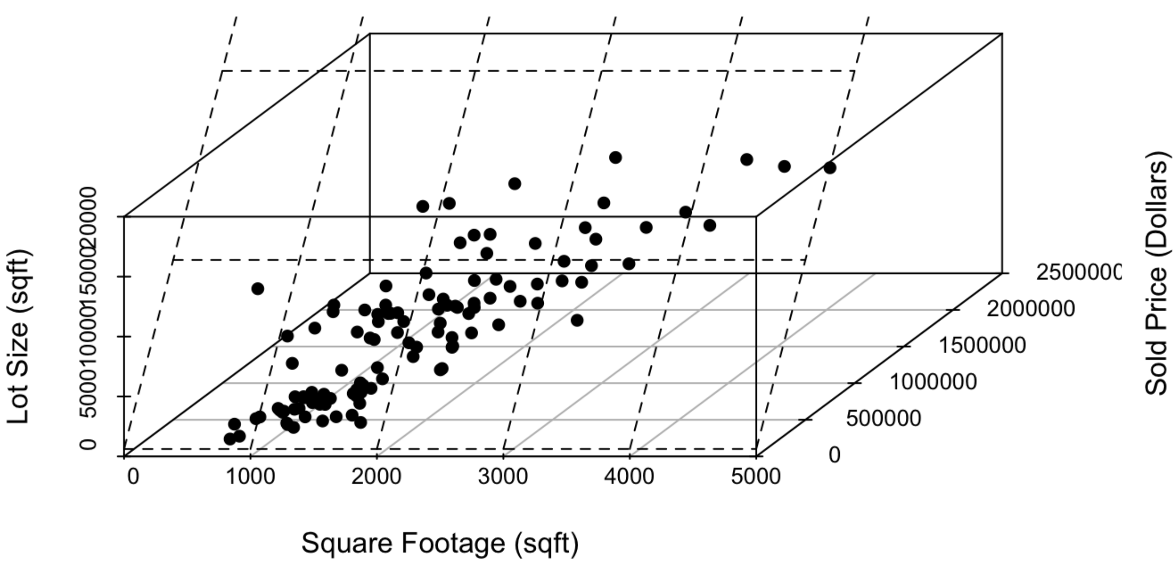


When looking at similarly generated distributions for the square footage of the lot (LotSize) in Figures 5.7 and 5.8 above, we found that the behavior is similar to the distribution of square footage of the home structure (Sqft) in Figures 5.5 and 5.6. This makes sense; from the correlation matrix in Figure 5.2, LotSize has a moderate to strong correlation to Sqft. Therefore, LotSize and Sqft may have high multicollinearity, or correlation, which may need to be addressed in the final model if both are included, as otherwise the statistical inferences may be less reliable.

D. 3D Scatterplot

As observed earlier in Figure 5.2, both LotSize and Sqft have positive associations with SoldPrice, but may have different joint behavior. To investigate the joint relationship between the sold price of the house (SoldPrice) and both LotSize and Sqft, we generated a 3D scatterplot depicting this relationship. This can be seen in Figure 5.9, below.

Figure 5.9: 3D Scatterplot of Sqft and LotSize as Predictors of SoldPrice



As seen by the plane fitted through the model, both LotSize and Sqft appear to have a joint positive relationship with SoldPrice. In other words, if either the lot size or square footage increase while the other is held constant, the selling price of the home increases as well. An interaction between LotSize and Sqft is likely not necessary, since the behavior of LotSize does not seem to change depending on Sqft, and vice versa.

Based on these visual observations, we proceeded to perform a cyclical process of variable processing, residual analysis, and model selection. The subsequent sections and their procedures should therefore not be seen as a sequential process, but three interworking parts of fitting a linear model.

VI. Variable Pre-Processing

Variable pre-processing was a cyclical process that went hand in hand with fitting a linear model, together involving three steps that were repeated in this sequence as many times as were necessary:

1. Variable selection using the best subsets procedure
2. Residual analysis of the resulting model
3. Variable processing and transformation

A. First Cycle

The first cycle of the process included using best subsets variable selection technique on all untransformed variables which examined all possible models. We then selected a few models to check model assumptions as well as added variable plots for the models and variables. The following variables were initially selected for the model:

Sqft
LotSize
HomeTypeNew

One major problem was discovered for the lot size variable. We noticed several observations in the dataset with extremely large lot sizes. We see extreme observations forcing a positive trend between lot size and sale price in the first added variable plot on the left. This point is considered extremely influential and mimics the influence plot from Anscombe's quartet, as seen in Figure 6.1. We decided these observations do not match the other homes and should be treated differently.

Figure 6.1: Added Variable Plot for LotSize, Before Subsetting

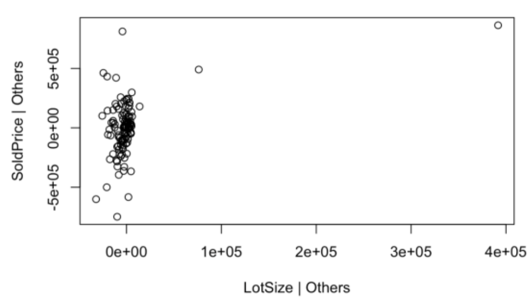
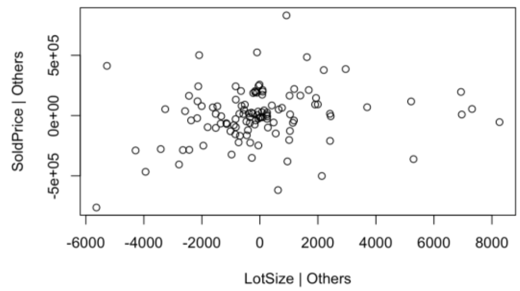


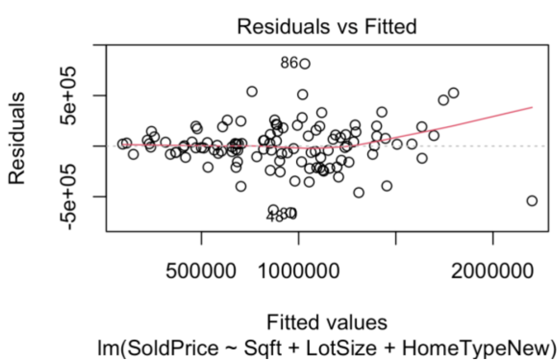
Figure 6.2: Added Variable Plot for LotSize, After Subsetting

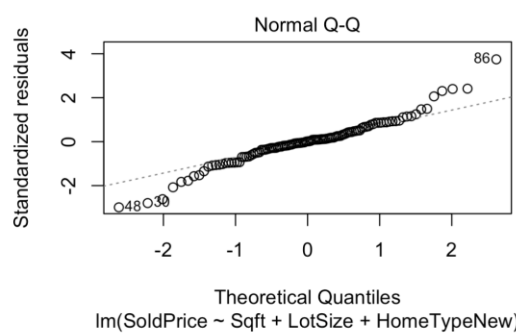


As a result, to avoid the possibility of outliers becoming influential on our model, we decided to permanently focus on properties that had a lot size below 16,000 square feet, thus cleaning off outlying LotSize observations. This was appropriate given observations in data visualization and variable pre-processing, seen in Figures 5.1, 5.2 and 6.1. After subsetting, we re-split the data randomly under seed 678 once more for validation resulting in 113 observations in the training data and twenty-eight observations in the test data to be used for validation. The result can be seen in the added variable plot in Figure 6.2.

The residual versus predicted value plots for the model demonstrated violations in equal variance and normality. Below we see a fanning pattern indicating unequal variance in the residual versus fitted value plot. We also see obvious deviations from the diagonal line in the normal quantile plot, indicating a violation of the normality assumption. These two violations indicate necessary transformations in the response variable as well as the predictors if necessary to fit regression assumptions.

Figure 6.3: Residual vs Fitted and Normal Quantile Plot, before Variable Processing

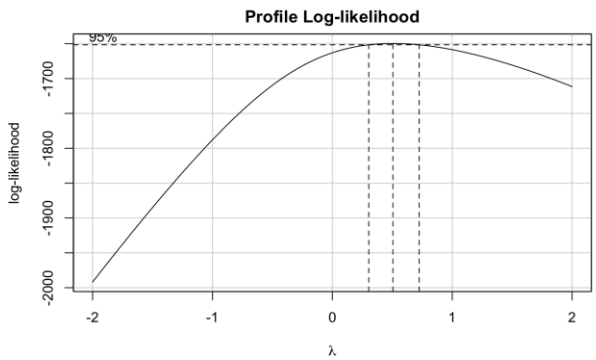




We then used the Box-Cox procedure to recommend a transformation of the response which minimized the sum of squared errors (SSE) of the model. As seen in the plot in Figure 6.4, Box-Cox procedure recommends decreasing the power of the response variable to around ½.

Therefore, we then repeated the best subsets procedure using all variables to predict the square root of sale price for the houses.

Figure 6.4: Box Cox Output for Recommended Transformation of Response Variable (SoldPrice)



B. Second Cycle

Using the best subsets procedure a second time, we were given similar results for best models and selected the same three variables. We could then test model assumptions for these models again to see the results of the transformation. The residual versus fitted plot depicted in the next section, in Figure 7.1, shows the equal variance violation was corrected by the transformation. Unfortunately, looking at the normal quantile plot, we still see slight deviations from the diagonal line indicating normality was not fully fixed and normality is still violated. The final step of variable pre-processing involved trying to adjust the model for the normality assumption with transformations of predictors. No transformation was found and thus we concluded this was the best we could do for the data at hand.

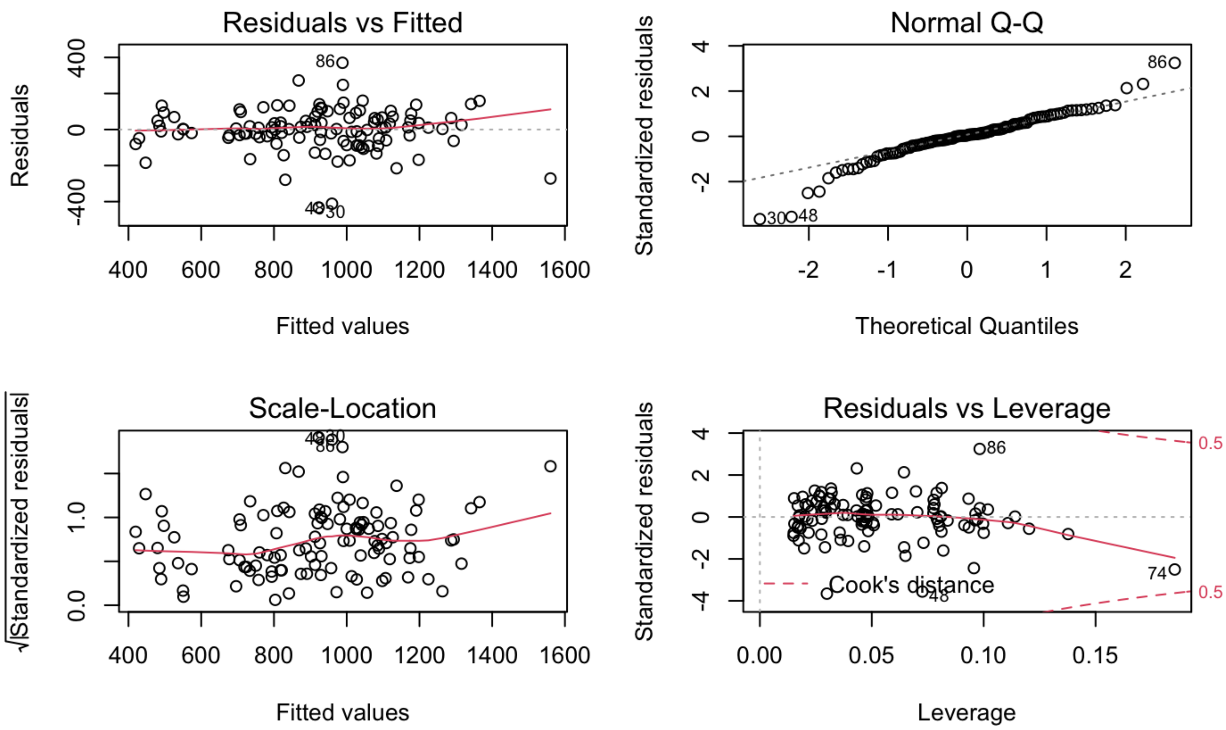
VII. Residual Analysis

To ensure that the model we selected fulfilled the basic assumptions for linear regression, we analyzed the plots found in Figures 7.1, 7.2 and 7.3.

After the cyclical process of variable pre-processing and model selection, the model we ended up with can be represented by the following:

$$\sqrt{SoldPrice} \sim Sqft + LotSize + HomeTypeNew$$

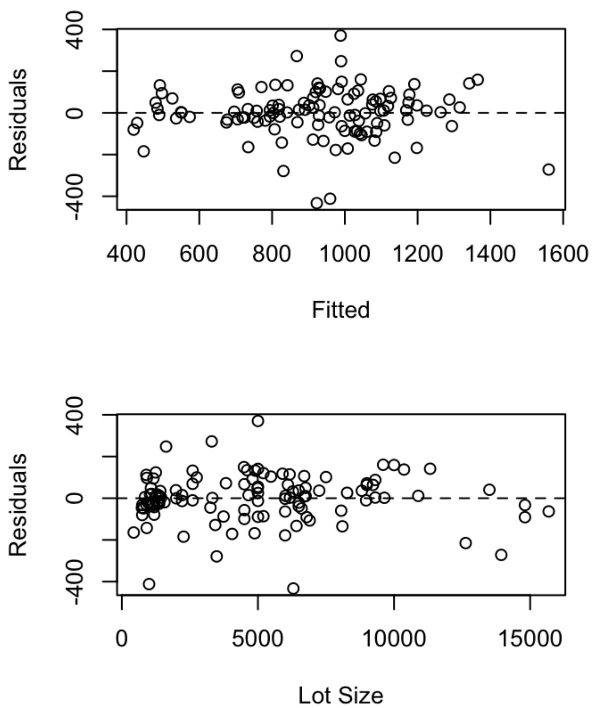
Figure 7.1: Four-in-One Model Diagnostic Plot



The four assumptions for linear regression include linearity, independence, normality of error, and homoscedasticity (equal variance of error). As an overall summary of plots that allow us to check for model assumptions, Figure 7.1 depicts a residual versus fitted value plot in the top left, a normal QQ plot in the top right, a scale-location plot in the bottom left, and a residual versus leverage plot in the bottom right.

From the residual versus fitted plot and the scale-location plot in Figure 7.1, it appears that the linearity and equal variance assumptions for the model are fulfilled, due to the lack of a fanning pattern or curvilinear pattern in the residuals. However, an inspection of the normal QQ plot indicates that there may be problems with normality. The residuals versus leverage plot indicates three outlying points, but no influential points, as all points appear to fall within half of Cook's distance. Points above 0.5 to 1 Cook's distance are considered influential and may disrupt model fitting.

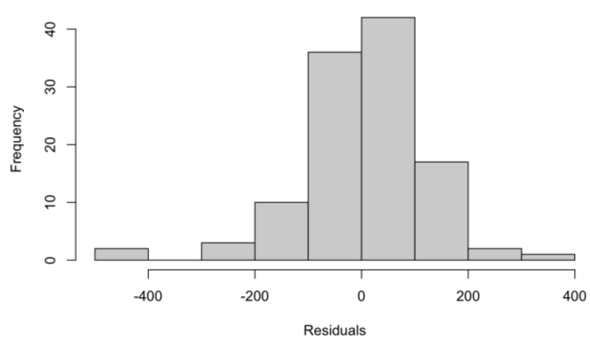
Figure 7.2: Residuals Plotted against Fitted Values, Sqft, LotSize, and HomeTypeNew



An inspection of the residuals plotted against each of the predictors in Figure 7.2 does not reveal any violations of the assumptions of equal variance or linearity.

After performing a Breusch-Pagan test for equality of variance, we concluded that there was not enough evidence that equality of variance was violated based on the p-value of 0.21.

Figure 7.3: Distribution of Model Residuals



A visual inspection of the distribution of residuals in Figure 7.3 appears to show normality of error. However, the Normal QQ plot does not support this assumption.

This can be further confirmed by the performance of a Shapiro-Wilk test, from which we concluded that there is strong evidence that the residuals are not normally distributed based on the p-value of 0.00025.

Due to the large size of the dataset, with 113 observations, and the lack of an appropriate transformation that would correct for normality of error, our team decided that we would overlook the normality assumption. However, this would likely result in a weaker predictive ability for the model. Therefore, we caution that any statistical inferences and predictions made from this model should be interpreted with this consideration in mind.

VIII. Fitting a Linear Model

The final model includes three variables with two quantitative variables and one categorical variable with four levels. Since the categorical variable has four levels, the equation includes three dummy variables for the different home types that shift the regression plane from the reference level, which is for condominiums.

$$\sqrt{SoldPrice} = 535.92 + (0.180)Sqft + (0.010)LotSize + (114.789)SingleFam - (259.680)Manufactured + (9.094)Other$$

The model does make sense contextually. In general, larger homes and larger properties tend to cost more. Additionally, different types of homes cost different amounts with single family homes considered to be more desirable than condominiums for many people.

For the training data, the model accounts for about 82% of the variation in the square root of the sale price of the homes. The typical prediction error, s, produced by the model is approximately 114 square root of dollars away from the least squares line.

As an interpretation of the intercept, the model estimates the square root of sale price for a home with a zero square foot interior, and a zero square foot lot size to be 535.9 square root dollars. However, this prediction is far outside the scope of our data and would be considered extrapolation.

We estimate that each additional increase of one hundred square feet in the interior size of the home increases the predicted square root of sale price by eighteen square root dollars after adjusting for the type of home as well as the lot size. The model also estimates the mean square root of sale price of single-family homes to be 114.8 square root dollars higher than the mean square root of sale price for condominiums. Comparing condominiums to manufactured homes, the mean square root of sale price for manufactured homes is estimated to be 259.7 square root dollars lower than condominiums. The model found no significant difference in the mean square root of sale price for homes in the “Other” category compared to condominiums after adjusting for the other predictors.

Using a Bonferroni-adjusted family-wise error rate of 0.05, we created confidence intervals for the parameters. We are 95% confident that the parameters in this model have values that fall within the intervals shown in Table 8.1.

Table 8.1: Confidence Intervals for Model Coefficients, 0.05 Family-Wise Error Rate

	Lower Bound	Upper Bound
β_{Sqft}	0.12	0.24
$\beta_{LotSize}$	-0.004	0.023
$\beta_{HomeTypeNewManufactured}$	-367.74	-146.77
$\beta_{HomeTypeNewOther}$	-80.81	157.54
$\beta_{HomeTypeNewSingleFam}$	17.73	224.61

There appear to be no significant issues with multicollinearity. The Generalized Variance Inflation Factor (GVIF), which measures multicollinearity values for each of the variables, are all below 5 as shown by Table 8.2, so we can safely conclude that the model has no issues with multicollinearity inflating the variances of the coefficients. This is somewhat surprising, given that there was a moderate association between Sqft and LotSize, as seen in Figure 5.2.

Table 8.2: Generalized Variance Inflation Factor Calculations for Final Model

Variable	GVIF
Sqft	1.63
LotSize	2.59
HomeTypeNew	2.12

IX. Statistical Inference

In order to see whether or not the combination of predictors chosen for the model are significant, we first performed an overall F-test on the model. As seen in Figure 9.1, we assumed for our null hypothesis that none of the coefficients for the predictors of sold price of the home are different from zero, and for our alternative hypothesis we stated that at least one predictor coefficient is significantly different from zero.

Figure 9.1: Overall F-Test for the Model

Model: $\sqrt{SoldPrice} \sim Sqft + LotSize + HomeTypeNew$

H0: $\beta_{Sqft} = \beta_{LotSize} = \beta_{SingleFam} = \beta_{Manufactured} = \beta_{Other} = 0$
HA: At least one $\beta_j \neq 0$

Test Statistic: F = 78.46 (on 5 and 107 degrees of freedom)
P-value < 0.001

Conclusion: Based on the small p-value, we reject the null hypothesis and conclude there is extremely strong evidence that at least one of the parameters is not zero.

As a result, we found sufficient evidence to conclude that the model we selected contains significantly useful predictors of sold price.

Next, we wanted to ensure that this model does not violate parsimony – that is, we want to ensure that there does not exist a simpler model with a single parameter that predicts the sold price of the home more efficiently. To this end, we performed a partial F-test on the model. As shown in Figure 5.2, the square footage of the home (Sqft) had a strong correlation with SoldPrice. Therefore, as seen in Figure 9.2, we compared our model to a model where Sqft is the sole predictor of SoldPrice.

Figure 9.2: Partial F-Test for the Model, Compared to Sqft as Sole Predictor

Full Model: $\sqrt{SoldPrice} \sim Sqft + LotSize + HomeTypeNew$
Reduced Model: $\sqrt{SoldPrice} \sim Sqft$

H0: $\beta_{LotSize} = \beta_{SingleFam} = \beta_{Manufactured} = \beta_{Other} = 0$
HA: At least one $\beta_j \neq 0$

Test Statistic: F = 29.898 (on 4 and 107 degrees of freedom)
P-value < 0.001

Conclusion: Based on the small p-value, we reject the null hypothesis and conclude there is extremely strong evidence that at least one of the parameters is not zero.

Based on the results seen in Figure 9.2, we found sufficient evidence to conclude that the size of the lot and type of home were still significant predictors after adjusting for square footage. In other words, our model was a better predictor of the square root of the sold price than when square footage was the only predictor.

We also noted in Figure 5.3 that there may be an interaction between the square footage (Sqft) and type of home (HomeTypeNew). In other words, the relationship between square footage and sold price of the home may change depending on the type of home. To see if this interaction was significant after including all other terms in the model, we again performed a partial F-test comparing the model with the interaction between Sqft and HomeTypeNew to our current model, the model without an interaction. This can be seen below in Figure 9.3.

Figure 9.3: Partial F-Test, Compared to Model with Sqft and HomeTypeNew Interaction

Full Model:

$$\sqrt{SoldPrice} \sim Sqft + LotSize + HomeTypeNew + Sqft{:}HomeTypeNew$$

Reduced Model: $\sqrt{SoldPrice} \sim Sqft + LotSize + HomeTypeNew$

H0: $\beta_{Sqft:SingleFam} = \beta_{Sqft:Manufactured} = \beta_{Sqft:Other} = 0$
HA: At least one $\beta_j \neq 0$

Test Statistic: F = 1.5758 (on 3 and 104 degrees of freedom)
P-value = 0.1997

Conclusion: Based on the large p-value, we fail to reject the null hypothesis and conclude there is not enough evidence that at least one of the parameters is not zero.

From the results seen in Figure 9.3, we found that after adjusting for the variables currently in our model, the interaction was not significant. In other words, after adjusting for square footage of the home, square footage of the lot, and type of home, there is not enough evidence that the relationship between square footage and price of the home depends on the type of the home.

Finally, as seen in Figure 5.9, we noted that there is likely not an interaction between LotSize and Sqft. In other words, the relationship between the sold price of the home and the square footage of the lot should not depend on the square footage of the home. Likewise, the relationship between the sold price of the home and the square footage of the home should not depend on that of the lot. To confirm this, we performed one more partial F-test comparing our model with the same model, with an additional interaction between LotSize and Sqft, as seen in Figure 9.4.

Figure 9.4: Partial F-Test, Compared to Model with Sqft and LotSize Interaction

Full Model: $\sqrt{SoldPrice} \sim Sqft + LotSize + HomeTypeNew + Sqft{:}LotSize$
Reduced Model: $\sqrt{SoldPrice} \sim Sqft + LotSize + HomeTypeNew$

H0: $\beta_{Sqft:LotSize} = 0$
HA: $\beta_{Sqft:LotSize} \neq 0$

Test Statistic: F = 2.0475 (on 1 and 106 degrees of freedom)

P-value = 0.1554

Conclusion: Based on the large p-value, we fail to reject the null hypothesis and conclude there is not enough evidence that the parameter is not equal to 0.

From the results seen in Figure 9.4, we confirmed our initial visual analysis and concluded that after adjusting for the existing model, LotSize does not depend on Sqft, so the interaction between the two is not a significant predictor and does not need to be included.

Using this model, we can create 95% confidence and prediction intervals for a home's selling price based on its square footage, lot size, and type of home. We decided to use the average values of square footage and lot size for a SLO-located single family home in our dataset, which are 1865 sq. ft. and 7079 sq. ft. respectively. The results can be found in Table 9.1.

Table 9.1: 95% Confidence and Prediction Intervals for Selling Price of Average SLO Home

	Lower Bound ($USD)	Upper Bound ($USD)
Confidence	1,061,410.94	1,184,762.59
Prediction	671,268.06	1,688,455.95

The values in the intervals have been back transformed from $\sqrt{\$USD}$ to $USD. As an interpretation of the intervals in Table 9.1, we are 95% confident that the average price for all single-family homes under 16,000 square feet in SLO sold in the 90 days before April 29th, 2022, with an average square footage (1865 sq. ft.) and average lot size (7079 sq. ft.) is between $1,061,411 and $1,184,763. Furthermore, we are 95% confident that an individual single-family home under 16,000 square feet with an average square footage (1865 sq. ft.) and average lot size (7079 sq. ft.) will have a sold price of between $671,268 and $1,688,455.95.

Once again, due to the violation of the assumption of normality of error, we advise caution in using the model to perform statistical inference and predictions.

X. Model Validation

Since we split the data in the beginning to create two separate data sets, one for training and the other for testing, we can test the model on data they were not directly modeled after to test its predictive capabilities by comparing model statistics to their predictive counterparts. If the two are similar, then it is plausible that the model has valid predictive ability. Looking at model performance for data outside the training data would be external validation.

A. External Validation

Using mean squared prediction error (MSPE), we can test the predictive ability of our model. MSPE mimics the calculation for mean square error (MSE) but uses the training model to predict outside observations and measures the mean of the residuals squared. Since the calculation of MSPE is like MSE, we can compare the two values to evaluate our model.

Table 10.1: External Validation Statistics

MSPE	7,739.8
MSE	13,680.2
$\sqrt{MSPE}$	87.98
$\sqrt{MSE} = s$	116.96

Comparing MSPE to MSE in Table 10.1, we see that MSPE and MSE appear to be vastly different in value, with MSE appearing to be almost double MSPE. Although on face value it would seem that the model is predicting external prices better than the prices in the training data, the vast difference in values indicates that the predictive ability of our model is likely not acceptable. Looking at the values for $\sqrt{MSPE}$ and the typical prediction error ($\sqrt{MSE}$ or s) in Table 10.1, we see that the typical prediction error for the test home selling prices is much smaller than the typical prediction error for the training home selling prices. The next step would be to check internal validation techniques to get other measures for the model performance since external validation may be influenced by the data split.

B. Internal Validation

Another way to measure the predictive abilities of the model would be to use the training homes for internal validation. The predicted residual estimates sum of squares (PRESS) calculates the prediction error predicting each observation using the model with that point removed. Summing up each of these values would produce a sum of squares error (SSE)-like statistic. This would then allow us to calculate an R-squared statistic for the predicted prices. R-squared is a measure of fit for the model, by measuring the proportion of variability in the selling prices of the homes explained by the model.

Table 10.2: Internal Validation Statistics

PRESS	1,767,794
SSE	1,545,867
$R^2{}_{pred}$	0.755
R^2	0.786

Table 10.2 shows promising results for the predictive ability for our model using a "leave-one-out" approach. We see the PRESS and SSE are not grossly different and the R^2 for training prices is similar to the R^2_{pred}, which predicts each observation using the model without the observation in the dataset. These statistics are promising and support that our model is doing a decent job predicting prices within the training data.

C. Full Model

Given that we have put the model to the test using internal and external validation, we can combine the two datasets to compute the model for the combined data. Looking at Table 10.3, we see that the coefficients of the model stay relatively the same for the full data compared to the training data model. We do see a larger difference in the mean square root of sale price for Other-type homes and condominiums for the full model. However, other than that slight difference, the coefficients across the models appear remarkably similar. This supports that the model does not "overfit" the data since the coefficients stay relatively constant with extra home observations added to the model.

Table 10.3 Model Coefficients for Training Data and Full Data

	Training Model	Full Data Model
Intercept	535.92	535.70
Sqft	0.18	0.18
LotSize	0.01	0.01
HomeTypeNewManufactured	-259.7	-257.3
HomeTypeNewOther	9.09	38.37
HomeTypeNewSingleFam	114.79	121.17

XI. Conclusion

Based on the model we fit to the selling prices of different SLO homes under 16,000 square feet sold within the 90 days before April 29th, 2022, the selling price of the home tends to depend on three major factors: the square footage of the house, the square footage of the property lot, and the type of home. There is compelling evidence that these variables are significant factors in determining how much the home sells for. Based on the coefficients, as the square footage of either the house or the lot increases, with all other variables in the model held constant, the selling price of the home tends to increase. Single family homes tend to be the most expensive, followed by “Other” (duplex, multi-family, townhouse, triplex, mobile manufactured, and planned development) homes, then condominiums, then manufactured homes, if square footage of the house and the lot are held constant.

However, our model is not perfect. As seen by the review of our linear regression assumptions, the model does not perfectly demonstrate normality of error, which means that the predictive power of the model is likely significantly weakened. This can be confirmed by our data validation, specifically our external model validation; it demonstrated that our model may not perfectly predict selling prices of our home based on external lot or footprint square footage and home type data. Furthermore, judging by the volatile behavior of home prices in the “Other” category, by aggregating multiple miscellaneous home types into an “Other” category for balance, we seem to weaken the model’s ability to predict home prices for “Other” type homes as a tradeoff. A stronger, more accurate investigation would involve subsetting more balanced data down to a specific type or types of home instead of aggregating disparate types into an “Other” category. Finally, our model is limited to a specific time and place – it is strongly likely that the model is only generalizable to the selling price of homes under 16,000 square feet in SLO, during the 90 days before April 29th, 2022. For all these reasons, we advise strong caution against using this model to predict the selling price of homes external to those used in this model, especially for different cities or periods of time.

Since our model predicts homes for such a specific time and place, future investigations of a similar nature may be of interest for other cities and other time periods. A way of verifying the strength of our model would be to see if searching for best models for selling prices in other time periods and cities using the same pool of variables results in similar predictive models that include lot size, home size, and type of home. Additionally, comparing these predictive linear models to predictions generated by existing home price appraisal algorithms, such as Zillow’s neural network-based model, Zestimate, may be a way to concretely gauge the predictive power of our model.

References

1. Qingqi Zhang, "Housing Price Prediction Based on Multiple Linear Regression", *Scientific Programming*, vol. 2021, Article ID 7678931, 9 pages, 2021. https://doi.org/10.1155/2021/7678931

2. Porcar Lahoz, Alicia E. “An Analysis of How Geographical Factors Affect Real Estate Prices.” *School of Architecture and the Built Environment Royal Institute of Technology*, Royal Institute of Technology, 2007, pp. 1–52.

Appendix

A. Data: SLO_Real_Estate_Feb-Apr_2022.xlsx
B. R Code: AppendixB_Rcode.rmd