Final Project

Module II: Modeling Data

Connor W. Benton

School of Data Science and Analytics, Kennesaw State University

DATA 1501: Introduction to Data Science

Dr. Priestley

July 14, 2022

Understand the Problem

With the information given, we know some key factors. The main issue is that the company is not maximizing their profits for two reasons. One reason is that the company will list the house on the market at a price that is too low. This is an issue because the company is not maximizing their profit; it is convenient that the listing leaves the market quickly, but more profit is possible. Another reason is that the company lists houses on the market at a price that is too high. This is an issue because the house is listed on the market for an extended period, which costs the company money.

What we have here are some listing inconsistencies; we need to find a way so the company can list all houses on the market at a suitable price. It would make sense to price houses based on the physical properties of the house, as well as the location, etc. With this in mind, a few pain points come to mind. First, it will be a bit difficult to gather and compile data on all listings this company has, it will take a good bit of resources to do this; however, since the outcomes will be more profit, one could argue it is a considerable tradeoff. Second, if we consider location of an area, that is an extremely broad statement – what makes a location good or bad? What type of location makes the price of a house higher or lower? We would need to keep ideas like this in mind as well.

We need to show this company what variables affect the listing price of a house, so they can correctly predict a reasonable listing price by finding trends and consistencies.

Problem Statement

Our current real estate listings on the market are not maximizing profits. We need to analyze how physical properties and key features affect the value of a property to accurately predict a listing price.

Diagnose the Data

We know that all data we work with is not in the proper format. Some values are missing, there are misspellings, issues due to human error, etc. In this case of our data, there are a good bit of errors, but nothing impossible from fixing. I would like to give a general overview of how the data was presented, then we will discuss what I did to fix those errors in the next section.

As for misspellings, first, looking at the data dictionary is necessary to properly ensure that the dictionary matched the data given. Looking at the LotConfig variable for lot configuration, we have three types: Inside, Corner, and CulDSac. When looking through the data, there were three typos that had a lot type of "CulSac" instead of "CulDSac." So, replacing these three values will be necessary for proper analysis. Second, we had an inconsistency with the data dictionary. For the BldgType variable, TwnhsI is an abbreviation for Townhouse Inside Unit and TwnhsE is an abbreviation for Townhouse End Unit; however, in the data, we have TwnhsE and Twnhs. So, replacing Twnhs with TwnhsI will be necessary to be consistent with the dictionary.

After looking through the data and seeing that misspellings and inconsistencies were dealt with, the next step is to analyze the number of missing values we have. For this data, there were a total of 18 missing values. We are missing 4 values for the YrSold variable, 3 values for the 2ndFlrSF variable, and 11 values for GrLivArea. We need to keep these missing values and mind and replace them accordingly, so we have a representative analysis.

Analyzing the shape of the variables is important to understand exactly what types of data we are working with. For a general overview of the shapes of variables, many were skewed right, with just a few symmetrical and skewed left variables. For more in-depth information on the specific variables and their shape, refer to the Diagnostics Table at the end of the section.

Finally, we need to identify outliers for each variable. For this paper, outliers were found by using the $1.5 * IQR$ rule. Almost every variable had outliers, few had extreme outliers, but one thing that comes to mind is that Sale Price had a total of 59 outliers. So, we should keep that in mind. Again, for more information on this, you may refer to the diagnostics table below.

Variable	Variable Categorization	Variable Type	Missing Values	Shape	Outliers	Notes
Id	Qualitative Categorical	N/A	0	N/A	N/A	None
LotConfig	Qualitative Categorical	Independent	0	302 Corner, 116 CulDSac, 1028 Inside	N/A	Typo: Replaced CulSac variable with CulDSac
BldgType	Qualitative Categorical	Independent	0	1210 1Fam, 29 2fmCon, 52 Duplex, 112 TwnhsE, 43 TwnhsI	N/A	To be consistent with the data dictionary, changed Twnhs to TwnhsI
ExterQual	Qualitative Ordinal	Independent	0	49 Ex, 481 Gd, 898 TA, 18 Fa, 0 Po	N/A	None
KitchenQual	Qualitative Ordinal	Independent	0	95 Ex, 582 Gd, 725 TA, 44 Fa, 0 Po	N/A	None
OverallQual	Qualitative Ordinal	Independent	0	Skewed Left	N/A	None
YearRemodAdd	Quantitative Continuous	Independent	0	Skewed Left, Bimodal	0	None
YrSold	Quantitative Continuous	Independent	4	Skewed Right	0	Replaced blanks with the median (2008)
1stFlrSF	Quantitative Continuous	Independent	0	Skewed Right	18	None
2ndFlrSF	Quantitative Continuous	Independent	3	Skewed Right	2	Replaced blanks with the median (0)
GrLivArea	Quantitative Continuous	Independent	11	Skewed Right	26	Replaced blanks with the median (1466)
BsmtFullBath	Quantitative Discrete	Independent	0	Skewed Right	1	None
FullBath	Quantitative Discrete	Independent	0	Roughly Symmetric	0	None
BedroomAbvGr	Quantitative Discrete	Independent	0	Roughly Symmetric	25	None
SalePrice	Quantitative Continuous	Dependent	0	Skewed Right	59	None

Clean the Data

To fix the spelling errors, we can just replace them with the appropriate variable. For the misspelling in the lot configuration variable, there are only three values that needed to be replaced. So, we can filter the data by "CulSac" and replace those values with "CulDSac" instead. Then, we have a consistent variable. As for the building type variable, we just change the variable "Twnhs" to "TwnhsI". In this case, there are 43 values to change, so filtering the data works just as well; however, a find and replace would also work fine. With those two changes, we already have more consistent data.

The next step is to deal with the missing values. As stated above, there are a total of 18 missing values, so it should not be too difficult. On top of that, all the missing values are quantitative variables, which means we can replace those missing values with the median to ensure the data is still representative and unbiased. So, the following calculations were done in Excel:

Median of YrSold: 2008

Median of 2ndFlrSF: 0

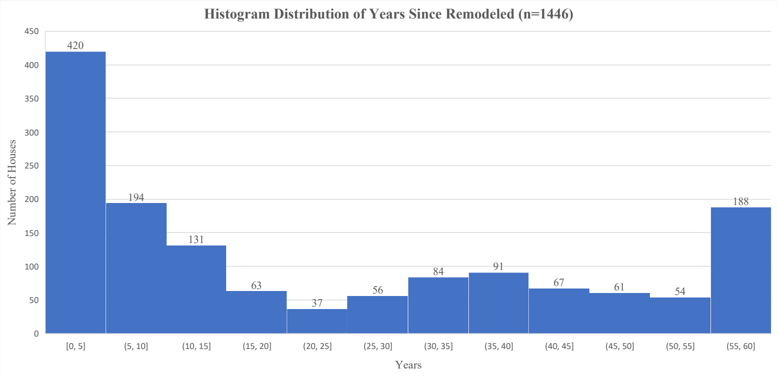
Median of GrLivArea: 1466

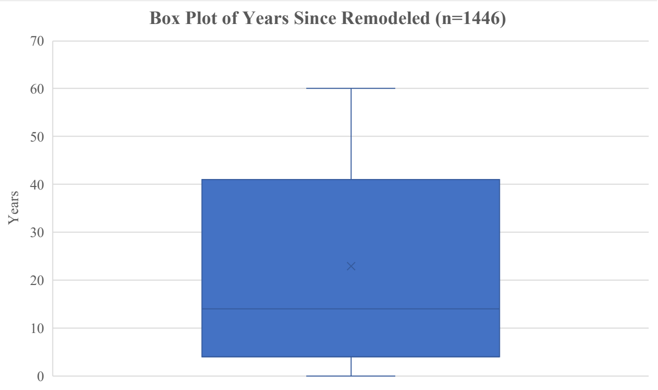
These values were substituted in for the missing values for their respective variables. We use the median since the median is not affected by outliers – this makes our data much more representative.

Univariate Quantitative Analysis

Variable	Minimum	Q1	Median	Q3	Maximum	Mean	Standard Deviation	IQR	95% Confidence Interval
YrSold	2006	2007	2008	2009	2010	2007.82	1.33	2	2007.75-2007.89
YrSinceRemod	0	4	14	41	60	23	20.63	37	21.94-24.06
1stFlrSF	334	882	1083.5	1383	3228	1156.15	370.26	501	1137.05-1175.25
2ndFlrSF	110	622.5	768	919	2065	800	272.2	296.5	778.55-821.45
GrLivArea	334	1131	1466	1771.75	4476	1507.26	500	640.75	1481.47-1533.05
BsmtFullBath	0	0	0	1	3	0.42	0.52	1	0.39-0.45
FullBath	0	1	2	2	3	1.56	0.55	1	1.53-1.59
BedroomAbvGr	0	2	3	3	6	2.86	0.8	1	2.82-2.9

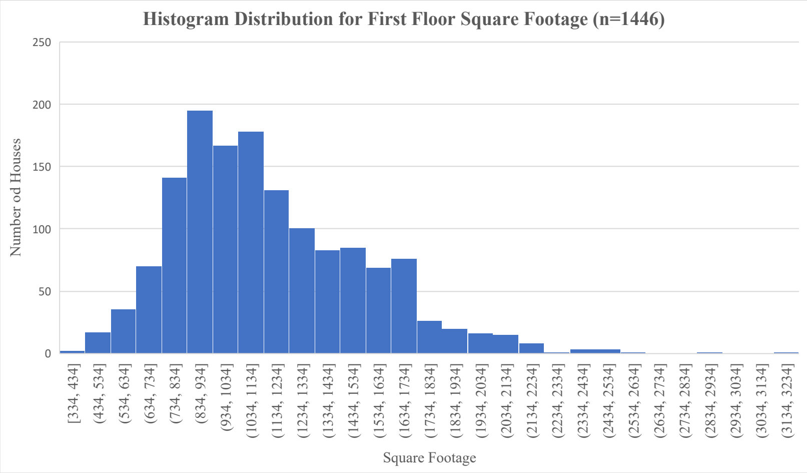
The table above depicts a generic summary of the topics we will cover in this section. We will go much more in-depth with each quantitative variable – to keep consistency, we will cover each variable in order of the above table. With that said, we will start with the year since remodeled variable since that would be more interesting to analyze than year sold.

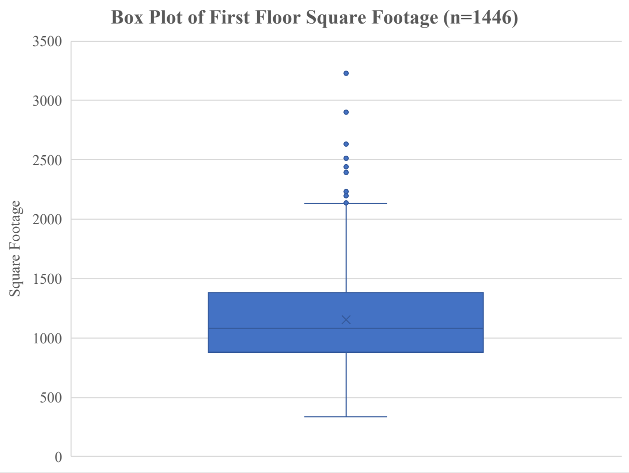




Looking at the above graphs, we have a histogram and a box plot of the number of years since the house had a remodel. To be more specific, this is the difference between the remodel date and the year sold. So, what we are really analyzing here is how quickly the house sold after it was remodeled. When looking at the histogram, we can see that it is skewed right, and one could argue it is bimodal with the two extremes on the ends. To back up this argument, we observe that the median is 14 years, while the mean is 23 years – this shows us that the mean is shifted to the right, so the distribution is skewed right.

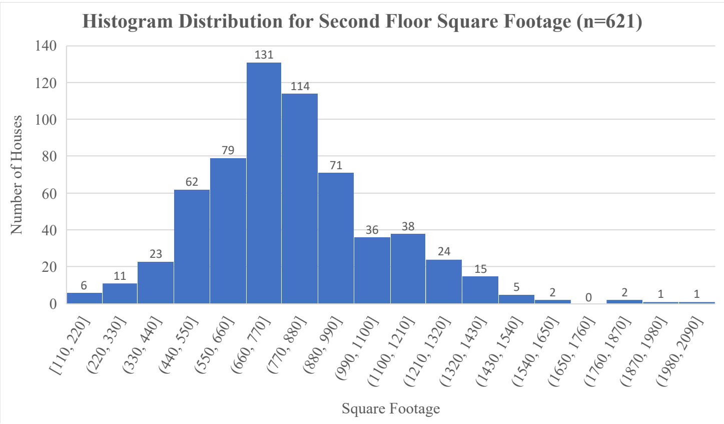
In this variable specifically, there were no outliers. Looking at the first quartile, we can see that 25% of the houses were sold, at most, 4 years after being remodeled. That tells us that 75% of the houses were sold more than 4 years after being remodeled. Looking at the IQR, the middle 50% of houses were sold between 4 years after being remodeled and 41 years after being remodeled, which is a large spread. Since the data is obviously skewed right, the appropriate measure of center would be the median and the appropriate measure of spread would be the IQR.

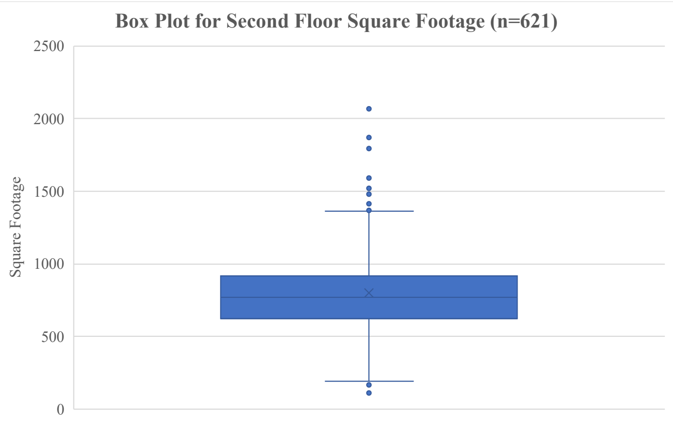




Here we have the distributions for first floor square footage. Looking at the histogram or the box plot, we can see that the variable has a unimodal, slightly right skewed distribution. If that is not obvious, looking at the mean for first floor square footage is about 1156 square feet, while the median is about 1084 square feet. So, since the mean is skewed higher than the median, implying that this is a right skewed distribution.

For outliers, there are a total of 18. Looking at the box plot, we see that the outliers mostly lie on the higher end. In fact, all the outliers lie on the higher end, which makes sense since there are some houses being sold that are exceptionally large. With these distributions, we see that the lower 25% of the houses have 882 square feet or less on the first floor. This tells us that 75% of the houses have more than 882 square feet on the first floor. Looking at the upper 25% of the houses, we see that those houses have at least 1383 square feet on the first floor. The middle 50% of houses have between 882 square feet and 1383 square feet on the first floor. Since this distribution is skewed to the right, a more reliable measurement of center would be the median. For spread, the interquartile range would be best.

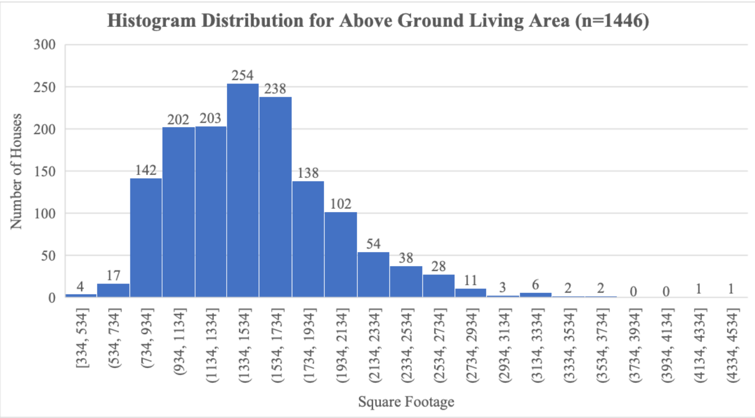


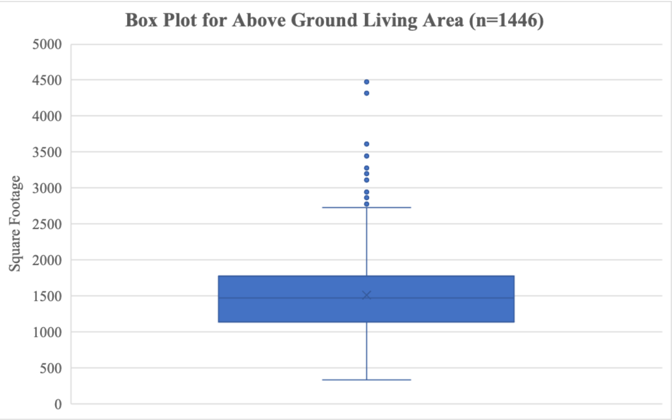


This distribution above is interesting to look at. Before diving into the analysis of this variable, we should discuss something. To make this data representative and not misleading, our sample size is much smaller, which sounds counterintuitive. Many houses in the dataset had no second floor, so the measurement for second floor square footage was 0. In fact, there were 825 houses in this dataset with no second floor. There would be no point in doing descriptive statistics for second floor square footage for houses without a second floor – it would be completely unrepresentative. So, these distributions only consider the 621 houses that have a second floor.

With that preface, we can begin looking at this data. From the table at the very beginning of this section, we see that the mean square footage for the second floor is 800 square feet, and the median is 768 square feet. With that said, we can see that the data is slightly skewed right since the mean is a bit larger than the median. Of course, we can also see that the data is unimodal.

After only considering the nonzero values of second floor square footage there are a total of 18 outliers. Most outliers lie on the higher end, but there are two outliers that are on the lower end. Even though these are outliers, it is still fine to have them since these values are plausible for houses. Looking at the box plot, the first quartile tells us that the lower 25% of houses have less than 623 square feet on the second floor. So, 75% of the houses have more than 623 square feet on the second floor. In fact, the upper 25% have at least 919 square feet on the second floor. So, the middle 50% of houses measured have between 623 square feet and 919 square feet. As stated earlier, the data is slightly skewed right, so the best measurement for center and spread would be the median and interquartile range, respectively.

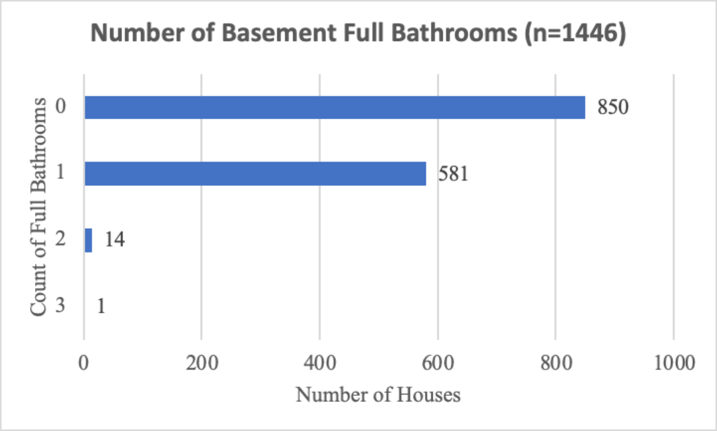


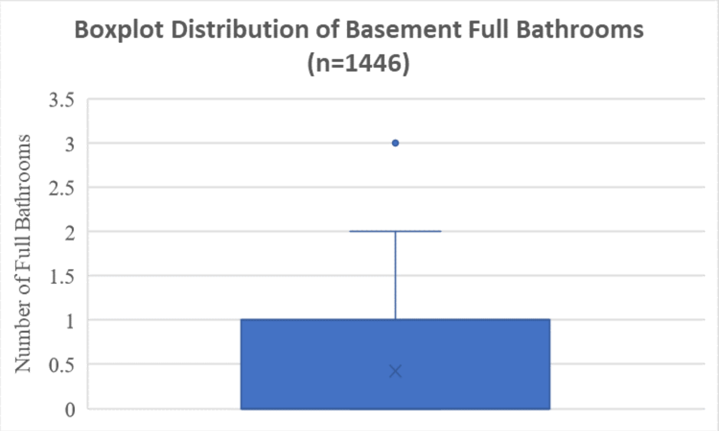


With the past few variables being analyzed and graphed visually, it seems that a trend is starting to develop here. Most of the variables we are working with are skewed right, which makes sense since we are talking about physical factors of houses. We should keep this in mind for the summary of our findings at the end of this paper. Let's discuss the above distributions. The above ground living area is basically the sum of the first floor and second floor square footage.

First, by looking at the data, we can see that our distribution is unimodal and skewed right. To further back that up, let's discuss the mean and median. Referencing the table at the beginning of this section, the mean above ground living area is about 1507 square feet while the median is 1466 square feet. Again, it seems like a trend is starting to show. Since the mean is larger than the median, we can be confident that the data is skewed to the right.

For this dataset specifically, the above ground living area has over 20 outliers. Since the first and second floor square footage had outliers, it would make sense that the above ground living area would have outliers since it is composed of the first and second floor square footage. These outliers are plausible, so there is no imputation required. The lower 25% of houses by above ground square footage have less than 1131 square feet. So, 75% of the houses have more than 1131 square feet. Looking at the middle 50% of the data, those houses have a square footage range between 1131 square feet and 1772 square feet. This also tells us that the upper 25% of houses by above ground square footage have more than 1772 square feet. Since the data is skewed to the right, the best measurement for center and spread would be the median and interquartile range, respectively.

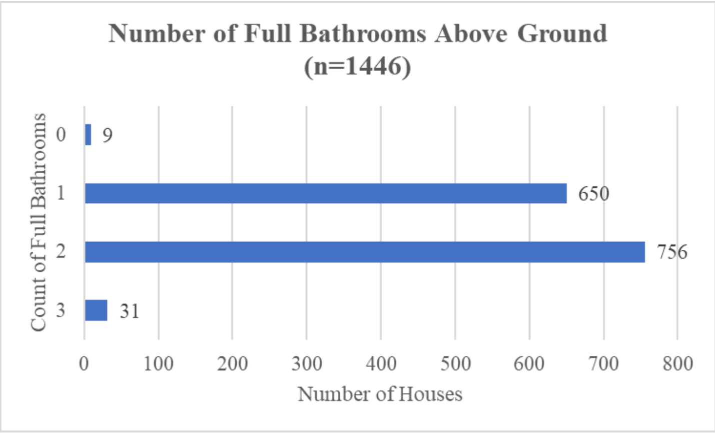


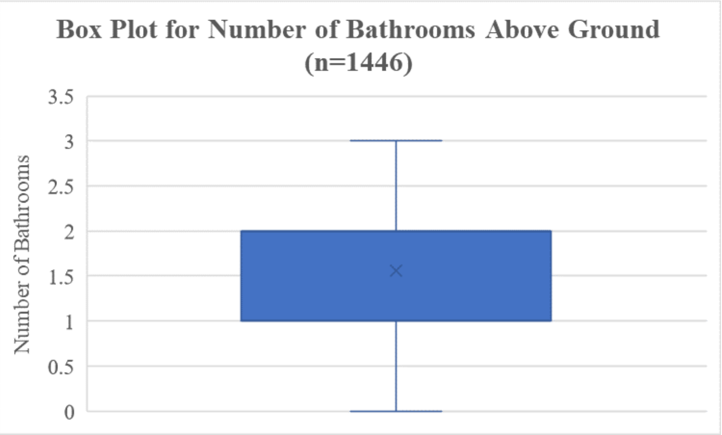


The above bar chart is a bit of a shift from the histograms and box plots we have been working with. This bar chart tells us the number of full bathrooms in the basement for each house. Of course, a histogram in this case would not make much sense. We can see that most of the houses in our dataset have no full bathrooms in the basement. This could be because there is no basement, or the basement has no bathroom. Since both options are possible, there is no reason to move the zero values for this analysis.

The median number of full bathrooms in the basement is 0, while the average is 0.42. With that said, we can see that the data is skewed right. To help visualize that better, we can look at the box plot above. On the upper end, we see that there is one outlier – a basement with 3 full bathrooms. In this case, there is no reason to impute the data since this is full plausible with the data we have. The first quartile of the data states that 25% of the houses have no basement full bathrooms, so 75% of the houses have more than 1 full bathroom in the basement. Looking at the middle 50% of the data, we see that 50% of the houses in the dataset have either 0 or 1 basement full bathrooms.

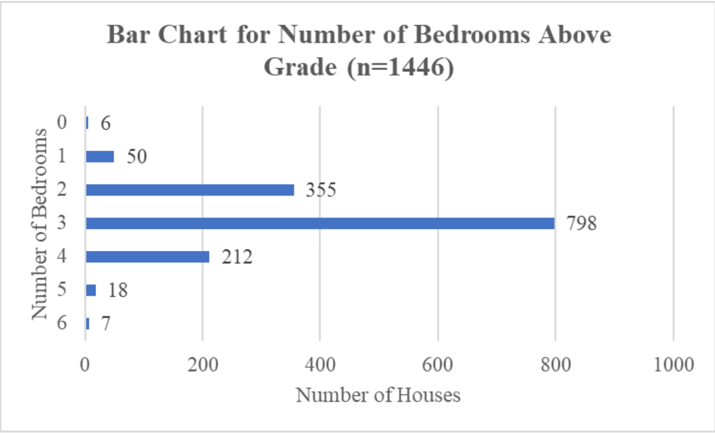
If we were to consider proper measurements for the center and spread, the median and interquartile range would be much better due to the slightly skewed nature of this data.

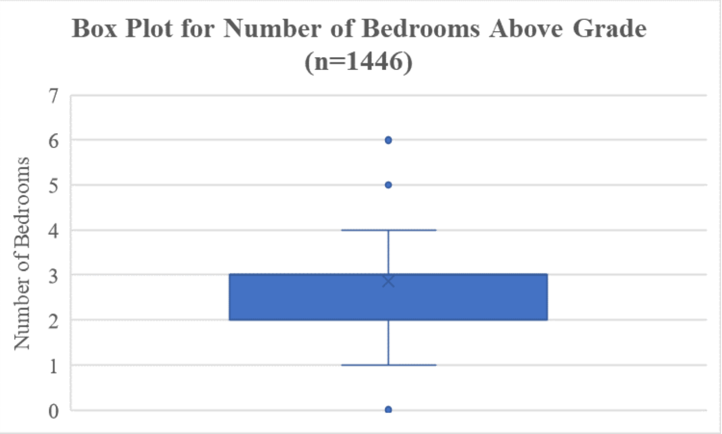




The analysis for this variable will be like the one above; however, we can see that the distribution is much different. First, this variable is the number of full bathrooms above ground, which most houses have. Some houses, however, only have half baths, which explains why there are some values of 0. Looking at the table at the beginning of this section, we see that the median number of full bathrooms above ground is 2, and the mean is 1.56. Using the $1.5 * IQR$ rule, it has been determined that there are no outliers for this variable. Considering the mean and median, this data is slightly skewed left since the mean is less than the median. Though the mean and median are not off by much, it probably would not matter which measurement of center and spread is used; however, just to be sure, using the median and interquartile range would be best.

Looking at the box plot, 25% of the houses in the dataset have either 0 or 1 full bathroom above ground. This tells us that 75% of the houses have 1-3 full bathrooms. The interquartile range suggests that 50% of the houses in the dataset have either 1 or 2 full bathrooms.

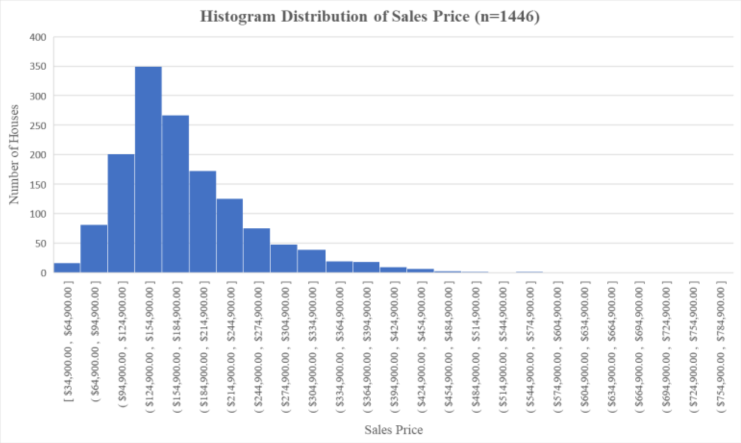


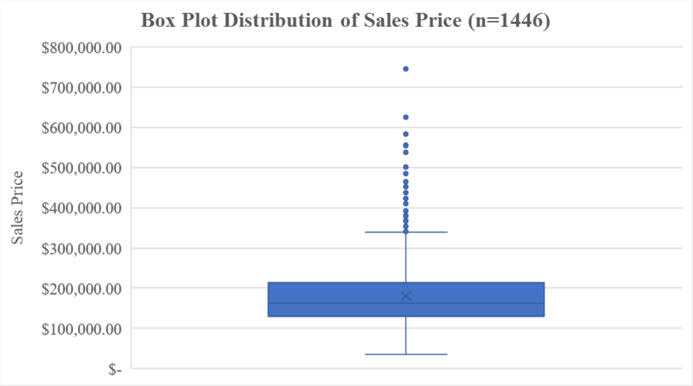


Here, we have a much more rich and spread-out distribution, unlike the other ones we have been working with. Again, since these measurements are discrete, a bar chart is much better for analysis here; however, just to visualize the spread, a box plot is included as well. We can see that this data is rather symmetric, which is interesting. The median number of bedrooms above grade is 3, while the mean is 2.86. These values are not that far from each other, so the measurement for center could be the mean or median. For the spread, since there are many outliers, the interquartile range may be best. Though, since these are discrete variables, the median and interquartile range would be much more interesting and contextual to analyze.

It may not be obvious, but there are a total of 26 outliers for this variable. Since these values are discrete, it is much more difficult to see them on the box plot. In this case, these outliers are plausible, so there is no reason to impute these values. Since we are using the median and interquartile range for this analysis, we will consider those values. The first quartile indicated that 25% of the houses in the dataset have at most 2 bedrooms above grade. So, 75% of the houses have 2 or more bedrooms above grade. The third quartile states that 25% of the houses have 3 or more bedrooms above grade. So, looking at the interquartile range, the middle 50% of houses have 2 or 3 bedrooms above grade.

Variable	Minimum	Q1	Median	Q3	Maximum	Mean	Standard Deviation	IQR	95% Confidence Interval
SalePrice	$34,900.00	$129,900.00	$162,950.00	$214,000.00	$755,000.00	$180,519.81	$ 78,652.10	$84,100.00	$176,462.50-$184,577.12



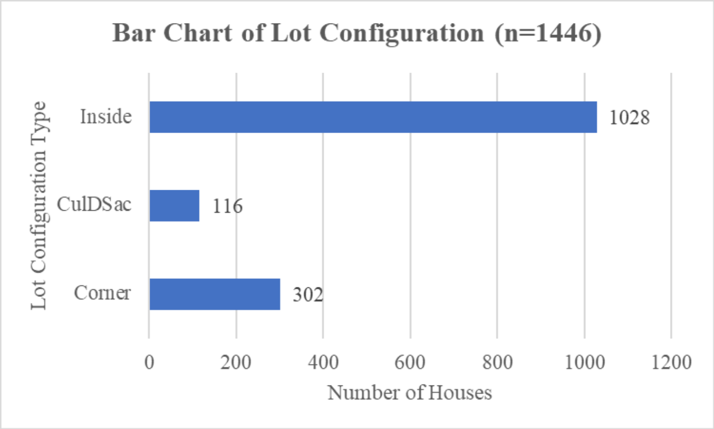


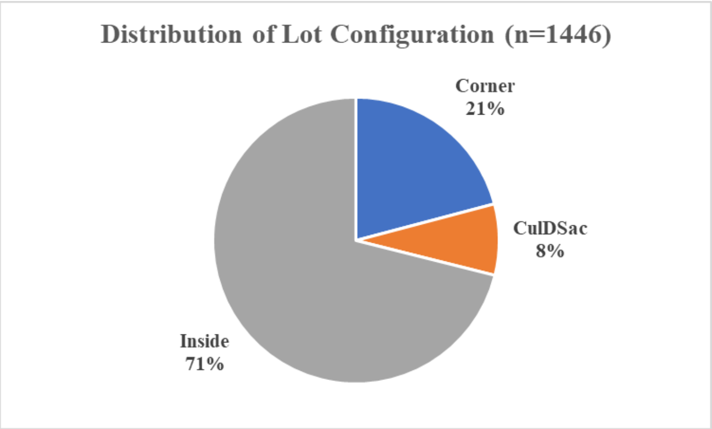
Here, we have the descriptive statistics as well as two distributions to visualize the sales price. This variable is the dependent variable, so it is important to understand exactly what is going on here. First, looking at the descriptive statistics table above, we see that the median sales price is $162,950, while the mean sales price is $180,519.81. This alone should indicate that the distribution of the sales price variable is heavily skewed right and affected by outliers. To be exact, there are a total of 59 outliers for the sales price, and all of them occur on the upper end. This makes sense, there are some houses that sell for larger prices; due to that, there is no reason to impute these outliers for now.

In this case, due to the multitude of outliers, the median is the best measurement for center. As for spread, again, due to the many outliers, the interquartile range is the best measurement. If we were to use the mean and standard deviation, our analysis may not be representative – the mean is heavily affected by outliers, so we need to tread carefully. By looking at the visual distribution above, as well as the median and mean, the sales price is unimodal and heavily skewed to the right.

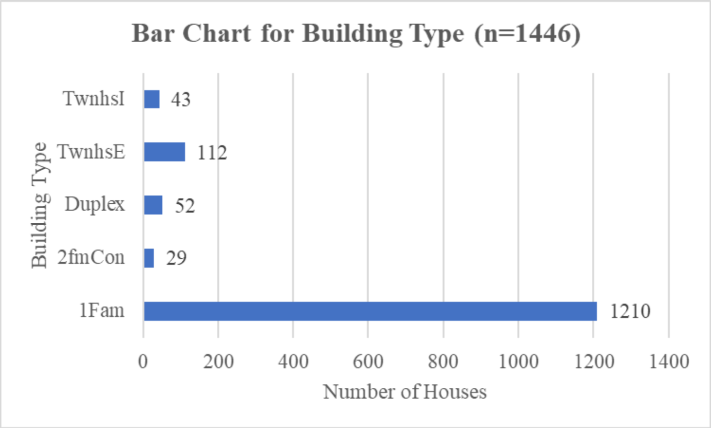
The first quartile states that 25% of the houses were listed for $129,900 or less, while 75% of the houses were listed for greater that value. The third quartile states that 75% of the houses were listed for $214,000 or less, while 25% of the houses were listed for greater that value. Looking at the middle 50% of the data, we see that 50% of the houses had a listing price between $129,900 and $214,000.

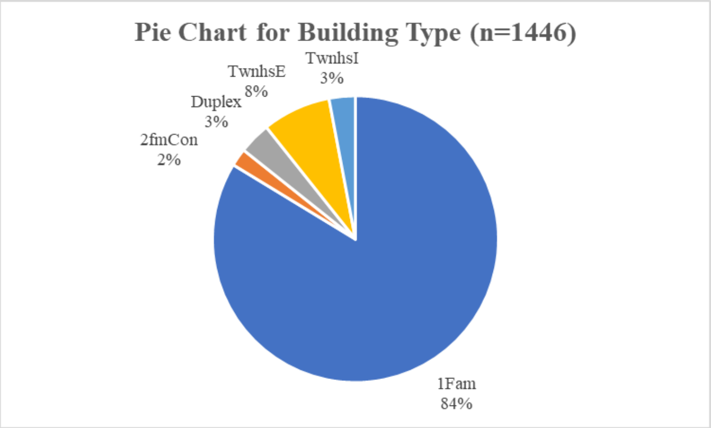
Univariate Qualitative Analysis



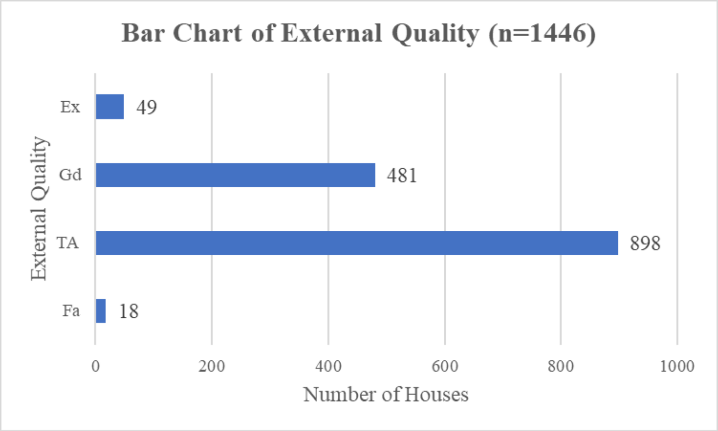


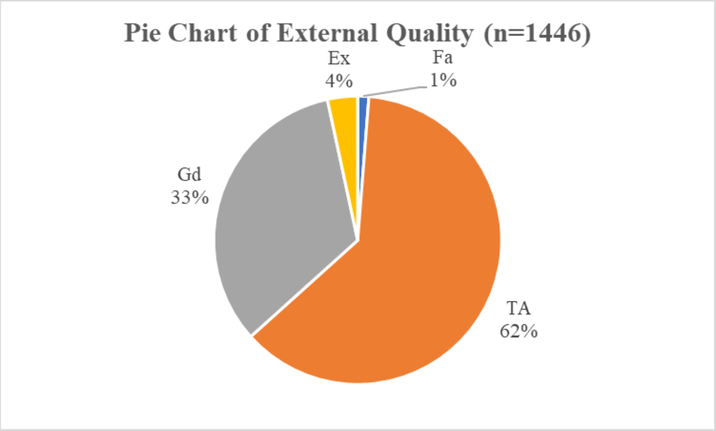
There are three types of lot configurations that a house may have: Inside Lot, Corner Lot, or a Cul-de-Sac. Looking at the distributions above, we see that most houses are inside lots. The second most frequent lot configuration is a corner lot, then the least occurring is a cul-de-sac. There are a total of 1446 houses in this sample, and 71% of the houses are inside lots, 21% are corner lots, and 8% are cul-de-sacs.



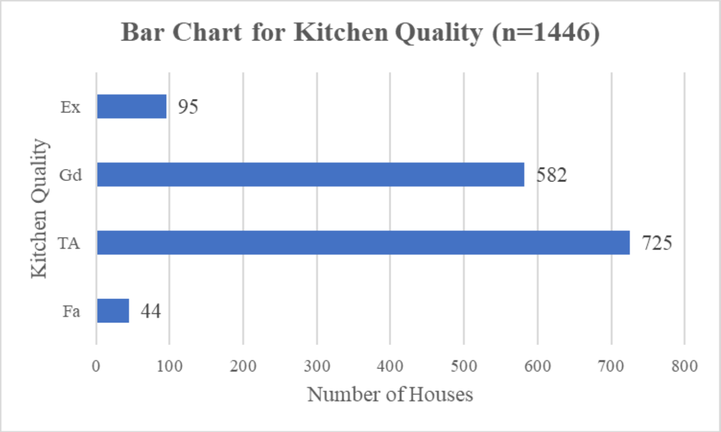


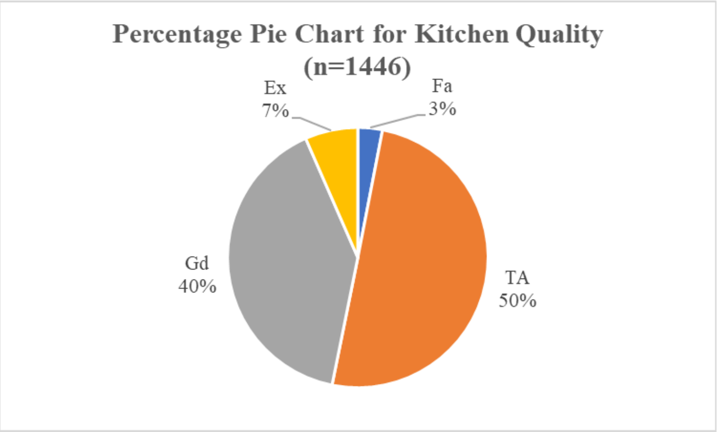
Like the lot configuration, it seems that there is a popular building type. There are five types of buildings that this sample has: Townhouse Inside Unit, Townhouse End Unit, Duplex, Two-family Conversion (duplex originally built as one-family dwelling), and Single-family detached. Referencing the bar chart above, most of the houses in the sample are single-family detached. This is the most popular building type, making up 84% of all houses in the sample. Other types following include TwnhsE (112), Duplex (52), TwnhsI (43), and 2fmCom (29).



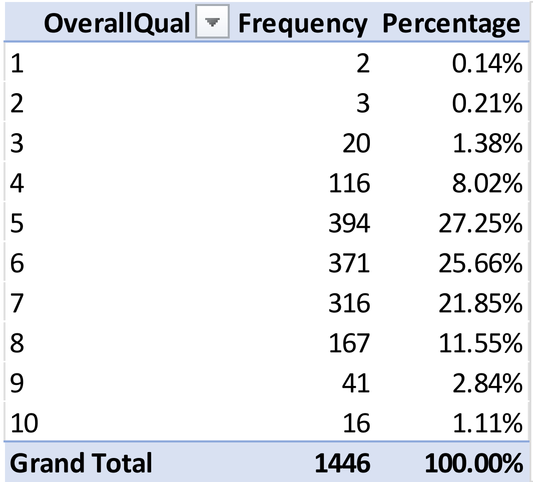


For context, the exterior quality of a house evaluates the quality of the material on the exterior. Unlike the earlier two variables, external quality of the house seems to be a bit more diverse. We see that an average/typical score is the most often occurring with almost 900 houses. Following behind is good quality with 481 houses. Behind that we have excellent quality with 49 houses and then a fair score with 18 houses. From the above pie chart, we can see that an average/typical score makes up 62% of all houses in the sample and a good score makes up about a third of all houses.

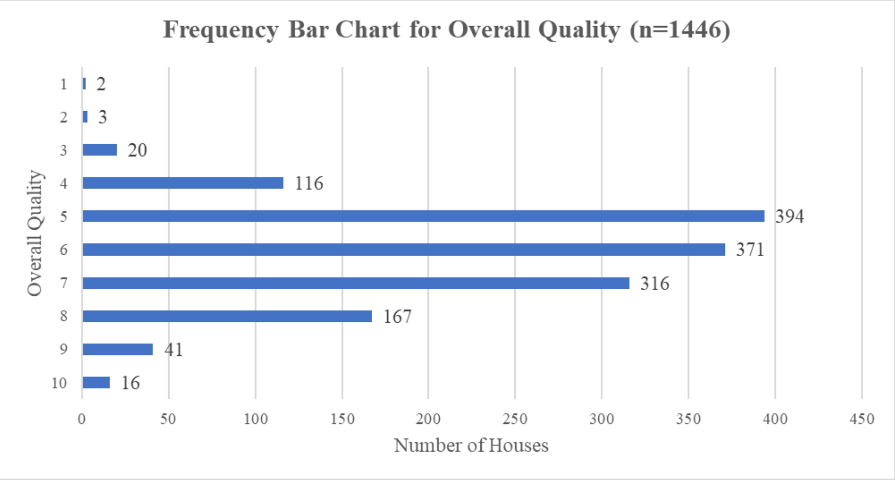




Another quality measurement in this sample is the kitchen quality. The rankings are the same as the exterior quality with excellent, good, average/typical, fair, and poor. To note, for the exterior quality and kitchen quality, there were no houses in the sample with a poor score. Referencing the above bar chart, an average/typical score was most occurring with 725 houses, followed closely by a good score, which had 582 observations. Then, there were 95 houses with an excellent score and 44 houses with a fair score for kitchen quality. In fact, an average/typical score made up 50% of houses, a good score made up 40% of houses, an excellent score made up 7% of houses, and a fair score made up 3% of houses.



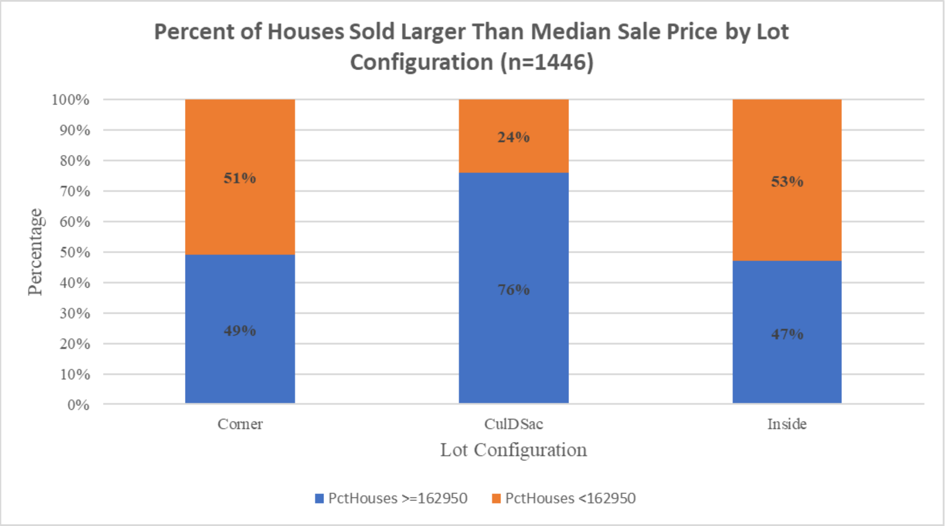
OverallQual	Frequency	Percentage
1	2	0.14%
2	3	0.21%
3	20	1.38%
4	116	8.02%
5	394	27.25%
6	371	25.66%
7	316	21.85%
8	167	11.55%
9	41	2.84%
10	16	1.11%
Grand Total	**1446**	**100.00%**



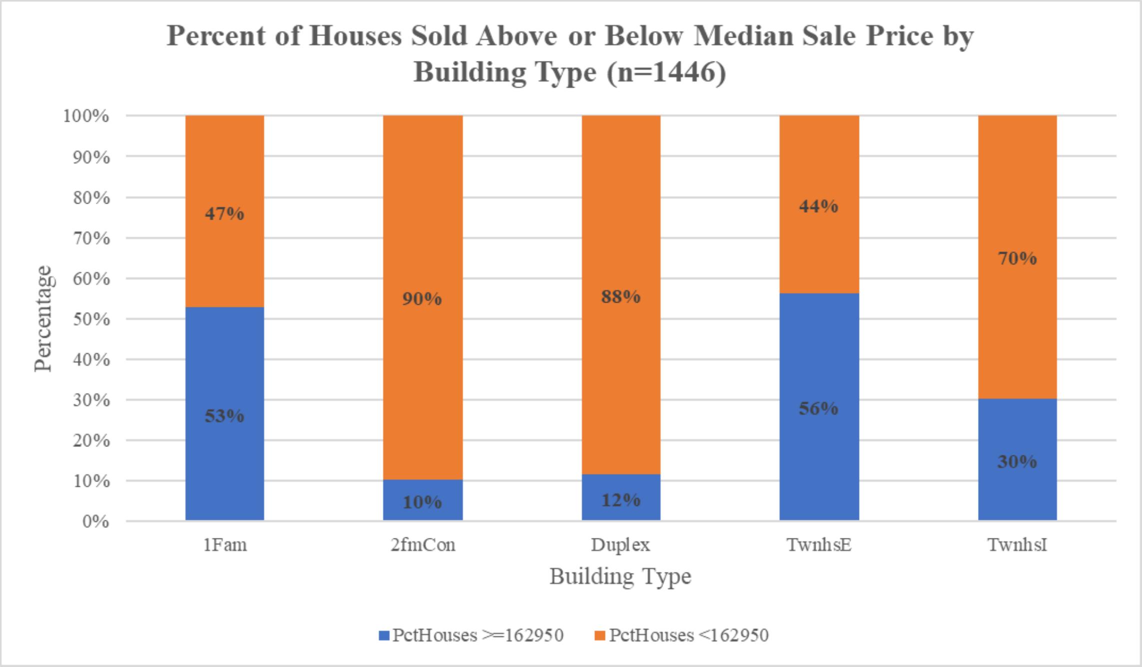
A frequency table and bar chart would be best to analyze the overall quality of the houses since a pie chart would be too cluttered. So, with the frequency table, percentages are still readily available. For overall quality, most of the values lie between 5 and 7, which make up about 74.76% of all houses in the sample. Towards the lower scores and higher scores, we see that the number of houses is much smaller.

Bivariate Analysis

Lot Configuration	PctHouses >=$162950	PctHouses <$162950
Corner	49%	51%
CulDSac	76%	24%
Inside	47%	53%



There's a lot going on here, so let's break down these graphics. A question that arose during the analysis is whether the lot configuration increases the value of a house or not. So, using a stacked column chart, we sort each lot configuration into two bins: sale price is greater than the median sale price ($162,950) or sale price is less than the median sale price. For Corner Lots and Inside Lots, there does not seem to be a trend; however, looking at the Cul-de-Sac is a different story. The Cul-de-Sac column tells us that 76% of Cul-de-Sac houses sold above the median sale price, while only 24% of Cul-de-Sac houses sold below the median sale price. This is rather interesting – it seems that a Cul-de-Sac Lot Configuration sells for much more than a Corner or Inside Lot. Of course, further analysis will be necessary to determine the significance of these results, but it is interesting to see how Cul-de-Sac lot configurations typically sell higher than the median sale price.

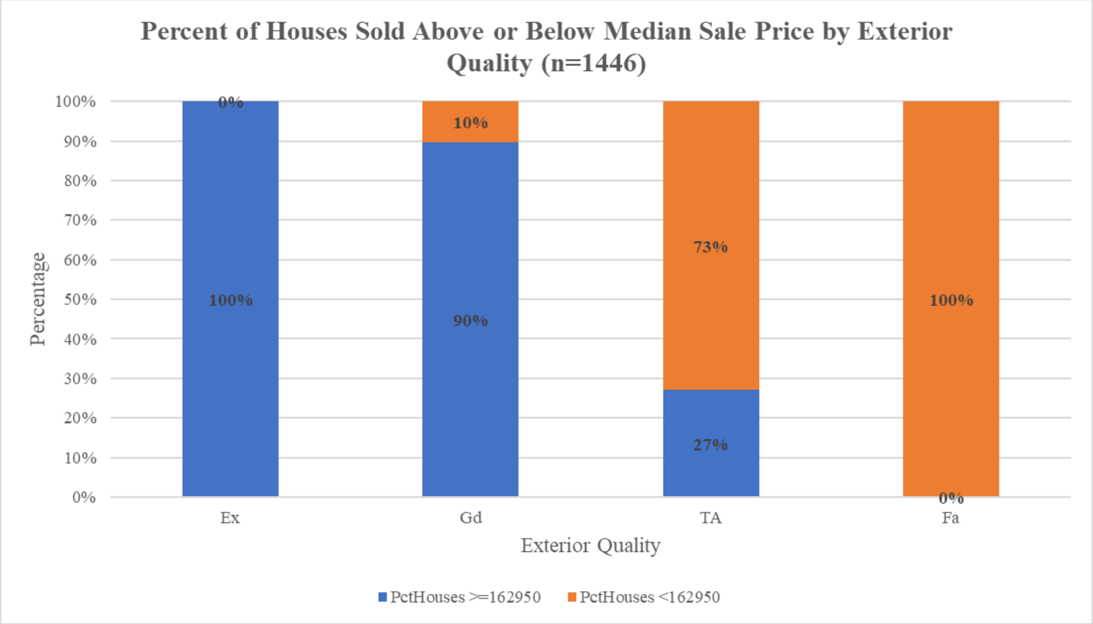


Building Type	PctHouses >=$162950	PctHouses <$162950
1Fam	53%	47%
2fmCon	10%	90%
Duplex	12%	88%
TwnhsE	56%	44%
TwnhsI	30%	70%

We should also use the same logic to see this relationship with other categorical variables. Another possible pattern could emerge with building type, and the findings are interesting. For reference, the median sale price is $162,950. Starting with single-family detached homes, 53% sold above the median sale price and 47% sold below the median sale price. For two-family conversion homes, 10% sold above the median sale price while 90% sold below the median sale price. Similarly, for duplex homes, 12% sold above the median sale price and 88% sold below the median sale price. For townhouse end units, 56% sold above the median sale price and 44% sold below the median sale price. Finally, for townhouse inside units, 30% sold above the median sale price and 70% sold below the median sale price.

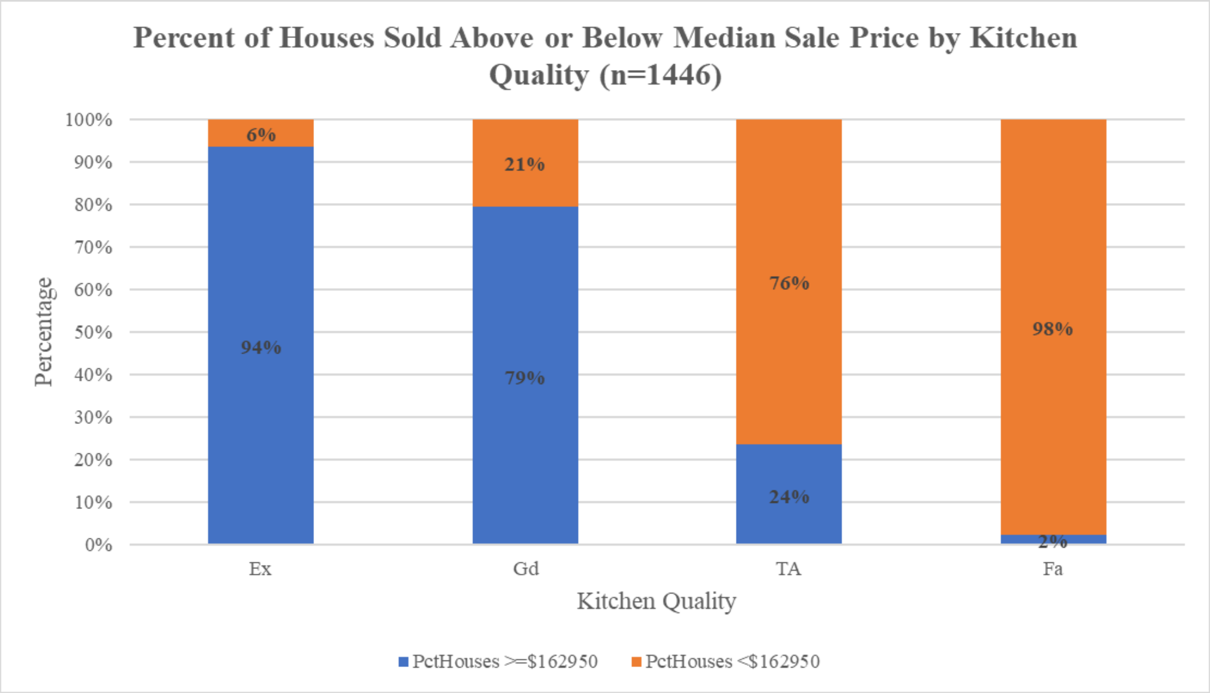
These findings are rather interesting – there seems to be a balance with single-family detached homes and townhouse end units; however, for two-family conversions and duplexes, the spread is quite large. It seems that two-family conversions, duplexes, and townhouse inside units sell at much lower prices compared to single-family detached and townhouse end units.

Exterior Quality	PctHouses >=$162950	PctHouses <$162950
Ex	100%	0%
Gd	90%	10%
TA	27%	73%
Fa	0%	100%



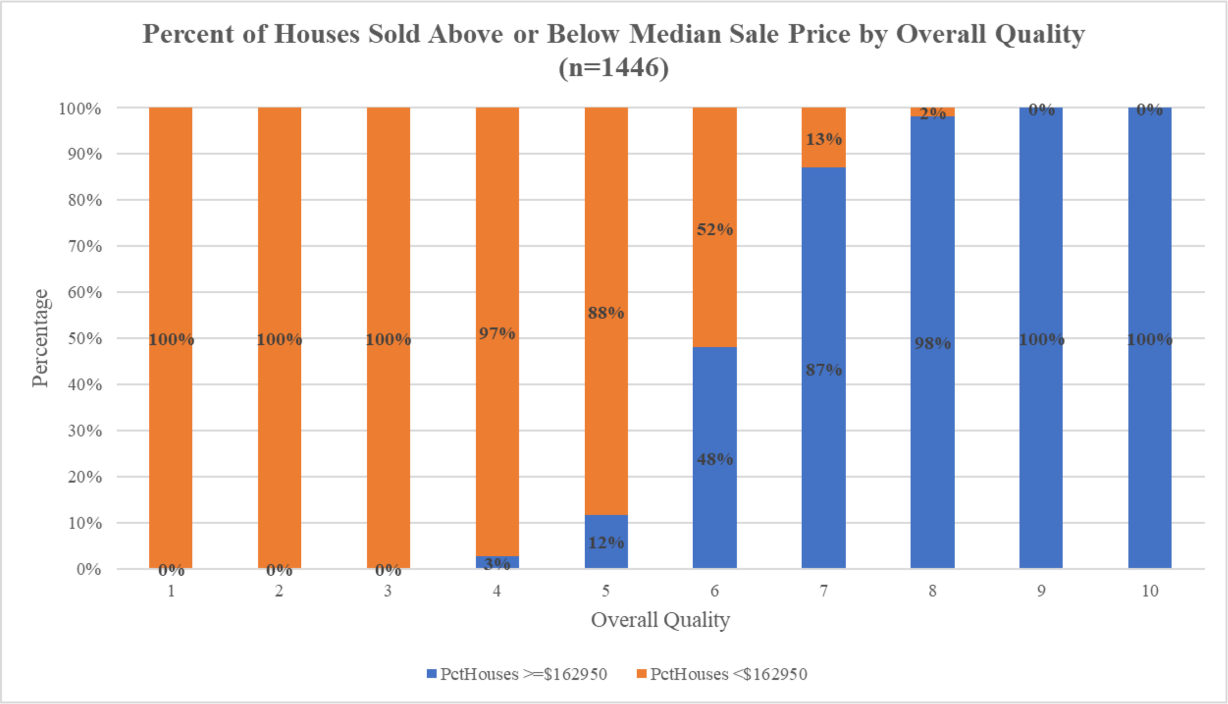
This graph here is probably the most informative analysis yet. For context, the exterior quality of a house is defined by the quality of the material on the exterior. For houses with an exterior quality score of excellent, 100% of those houses sold above the median sale price, which is $162,950. For houses with an exterior quality score of good, 90% of the houses sold above the median sale price, while 10% sold below the median sale price. For houses with an exterior quality score of typical/average, 27% of houses sold above the median sale price while 73% sold below the median sale price. Finally, houses with an exterior quality score of fair, 100% of the houses sold below the median sale price.

Kitchen Quality	PctHouses >=$162950	PctHouses <$162950
Ex	94%	6%
Gd	79%	21%
TA	24%	76%
Fa	2%	98%

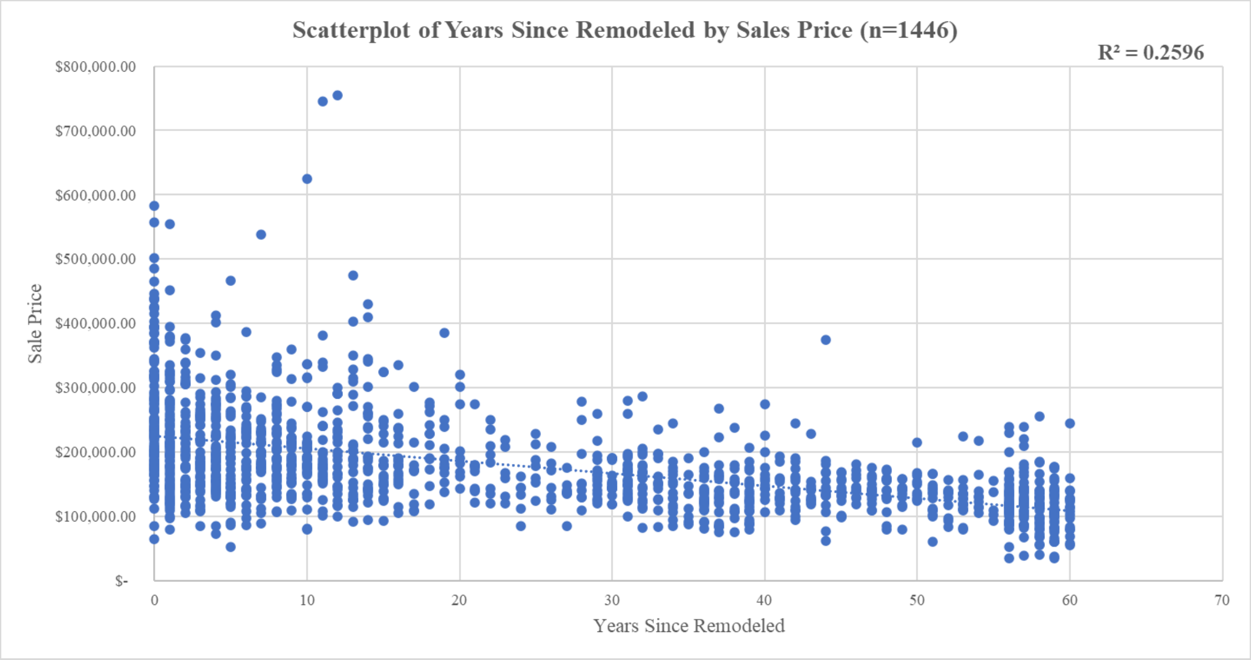


Like the exterior quality of a house, we see that kitchen quality seems to also have an impact on the sale price. Here, we can see that out of all houses that had a kitchen quality of excellent, 94% of them sold above the median sale price, which is $162,950. 6% sold below the median sale price. For the houses that had a kitchen quality score of good, 79% sold above the median sale price, while 21% sold below the median sale price. For houses that had a kitchen quality score of typical/average, 24% sold above the median sale price while 76% sold below the median sale price. Finally, for houses that had a kitchen quality score of fair, 2% sold above the median sale price while 98% sold below the median sale price.

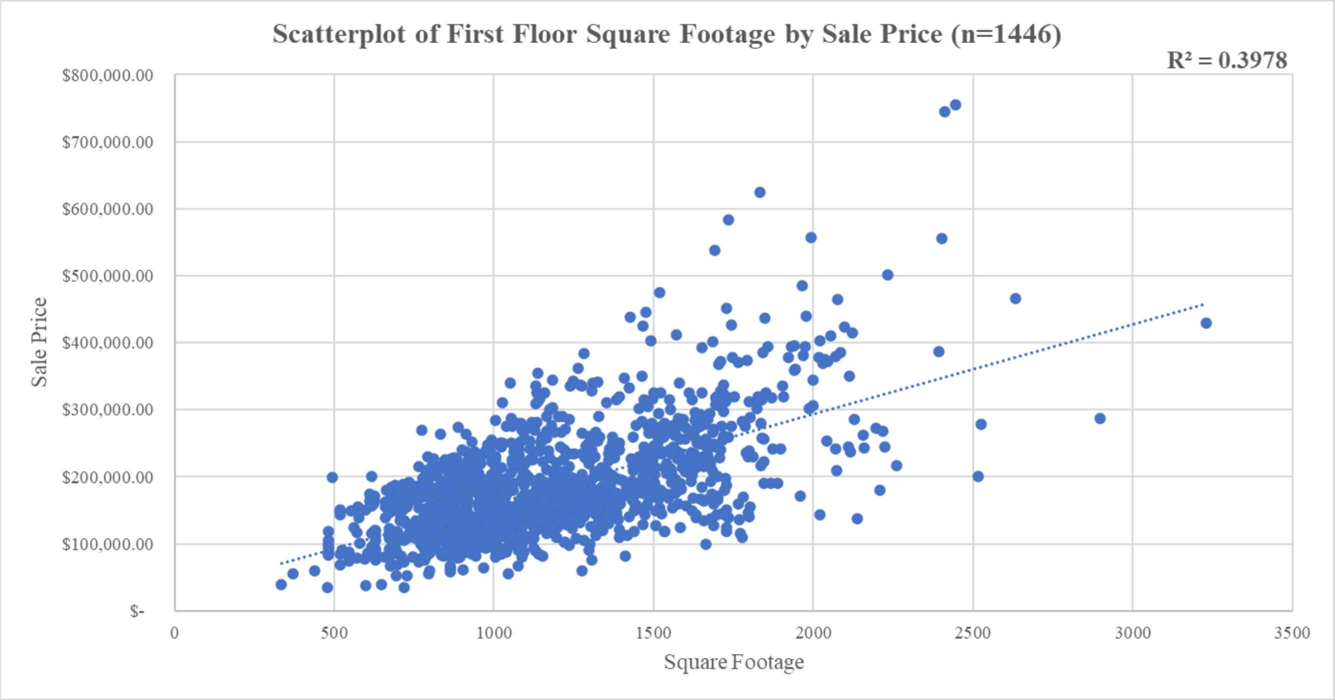
Overall Quality	PctHouses >=$162950	PctHouses <$162950
1	0%	100%
2	0%	100%
3	0%	100%
4	3%	97%
5	12%	88%
6	48%	52%
7	87%	13%
8	98%	2%
9	100%	0%
10	100%	0%



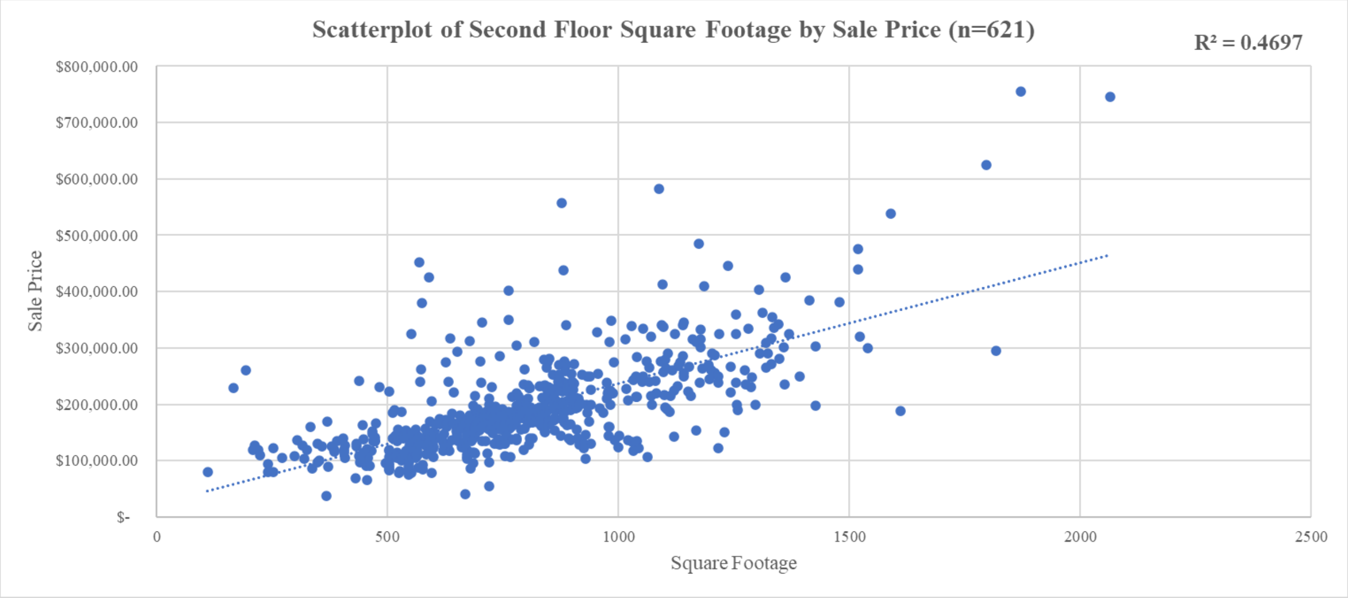
Finally, we have the relationship with overall quality. There does appear to be a strong correlation between the overall quality of a house and the sale price. As we can see from the above graph, the higher the overall quality score, the more houses sold above the median sale price. Houses that had an overall quality score of 1-3 all sold below the median sale price. We then start to see a shift when the overall quality score hits 4 – from all houses that sold with an overall quality score of 4, 3% sold above the median sale price and 97% sold below the median sale price. For houses with an overall quality score of 5, 12% sold above the median sale price and 88% sold below the median sale price. After that, we see a large jump; for houses with an overall quality score of 6, 48% of those houses sold above the median sale price while 52% sold below the median sale price. For houses with an overall quality of 7, 87% sold above the median sale price and 13% sold below the median sale price. For houses with an overall quality score of 8 or higher, only 2% of those houses sold under the median sale price, while the rest sold above the median sale price.



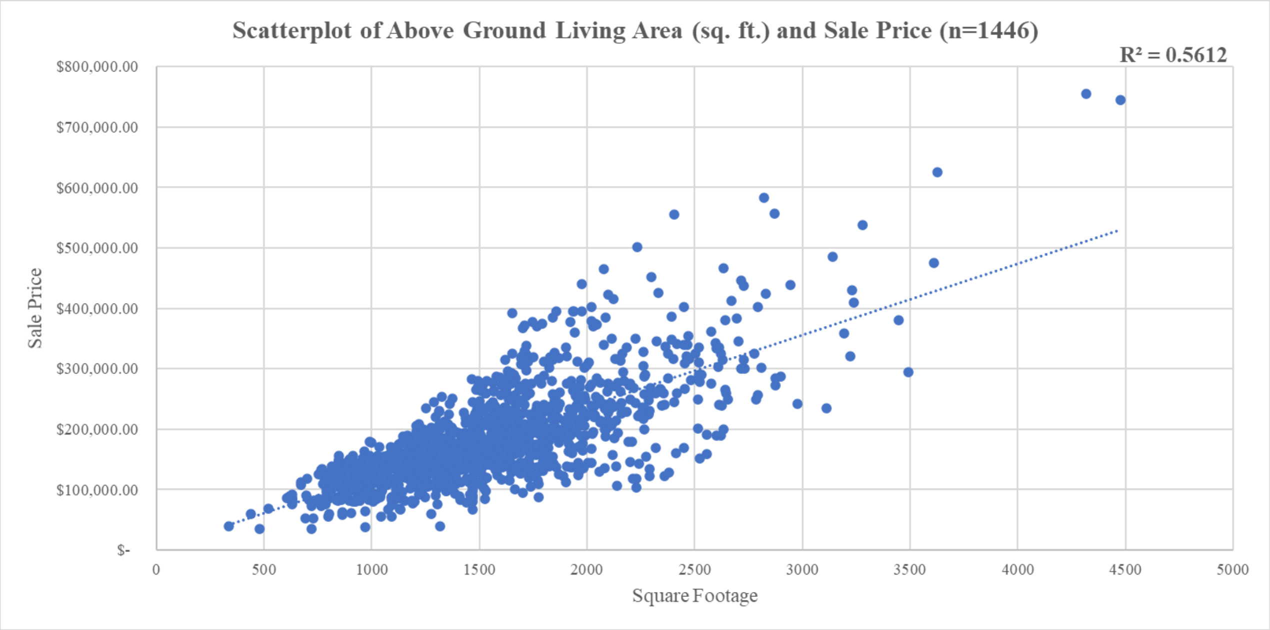
Looking at more quantitative relationships, we begin with looking at a scatterplot of years since remodeled by sales price. There seems to be a *slight* negative correlation, which is not bad – in fact, it tells us a lot. We can see that as the years since remodeled value increases, the sale price decreases. So, houses that have not been remodeled in a while tend to sell for less. However, looking at the trend, the data seems to stagger toward the 40-year mark, so much after that mark does not seem to affect as much. When looking at houses that have been recently remodeled, their sale price tends to be larger. More analysis would be needed to see if this finding is significant, mainly since the correlation is very slight.



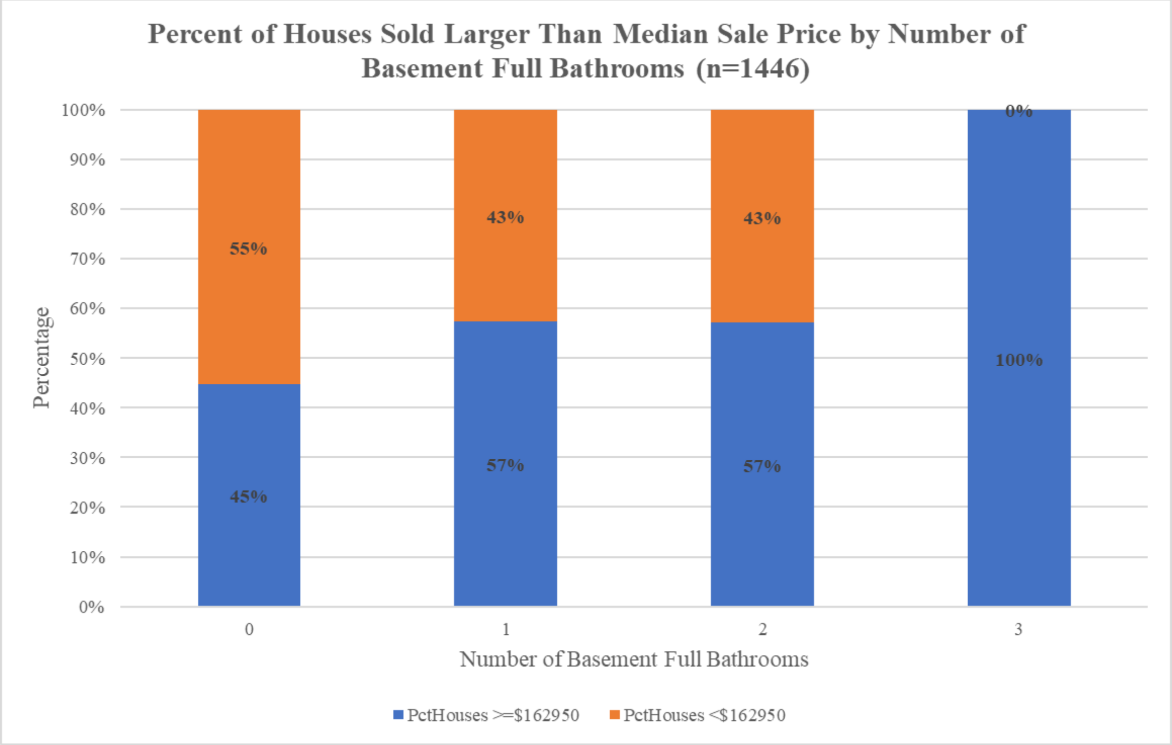
Next, we analyze the relationship between first floor square footage by sale price. We can see from the above scatterplot that there is a relatively strong, positive correlation between the first-floor square footage and sale price of a house. We can confidently say that as the square footage of a house is larger, the sale price increases. It is also worth noting how there is a large cluster from 500 sq. ft. to 1500 sq. ft. After that, the data points become more significantly scattered.



Note that the sample size for second floor square footage is 621 houses instead of 1446. This is because we are only considering houses with a nonzero square footage value for the second floor. As said earlier in the paper, there is no point to include houses without a second floor in this analysis. Generally, as the square footage if the second-floor increases, the sale price of the home also increased. We should also note the concentration between the 500 sq. ft. and 1000 sq. ft. values. After 1000 sq. ft., the datapoints begin to scatter outward.



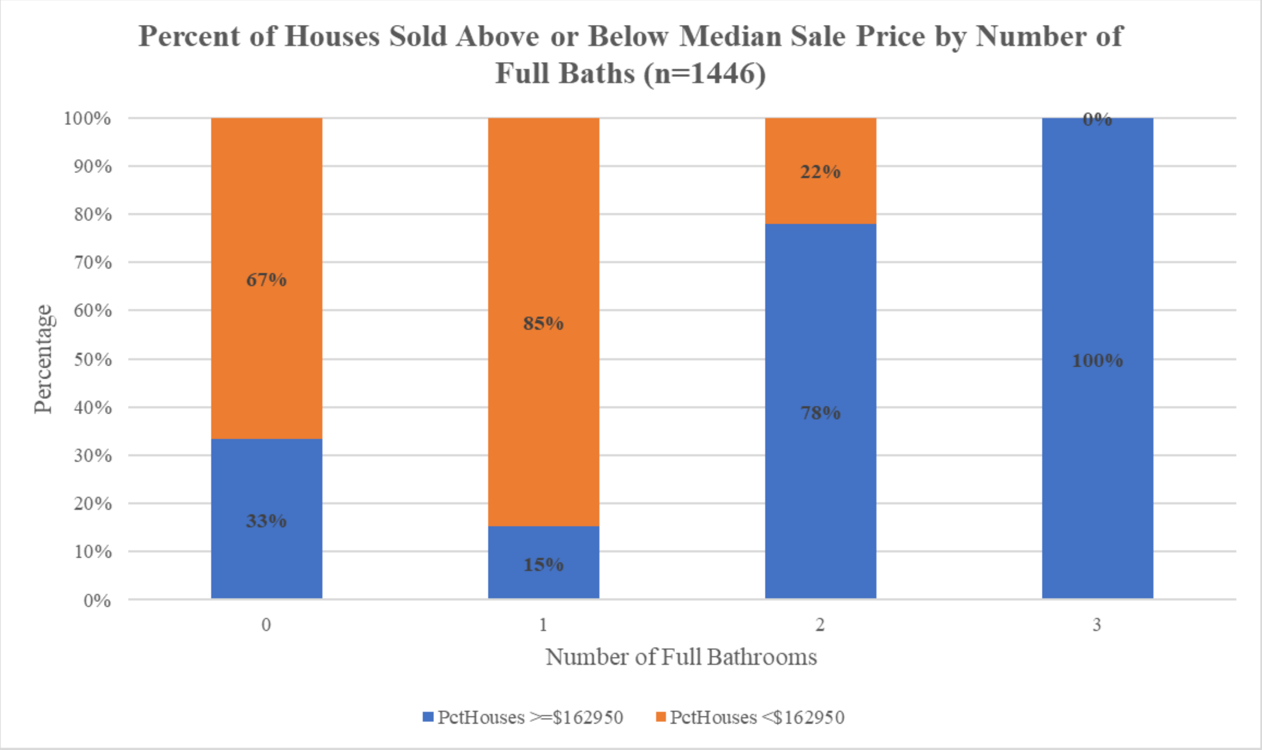
The final scatterplot we will be looking at will be the relationship between above ground living area and sale price. Again, there seems to be a strong, positive correlation between both variables. As the square footage increases, the sales price increases. We should also note the high concentration between 750 square feet and 2000 square feet. After 2000 square feet, the datapoints spread out more. This makes sense since there was a similar concentration when analyzing the scatterplots for first and second floor square footage by sale price.



NumBsmtBath	PctHouses >=$162950	PctHouses <$162950	Number of Houses
0	45%	55%	850
1	57%	43%	581
2	57%	43%	14
3	100%	0%	1

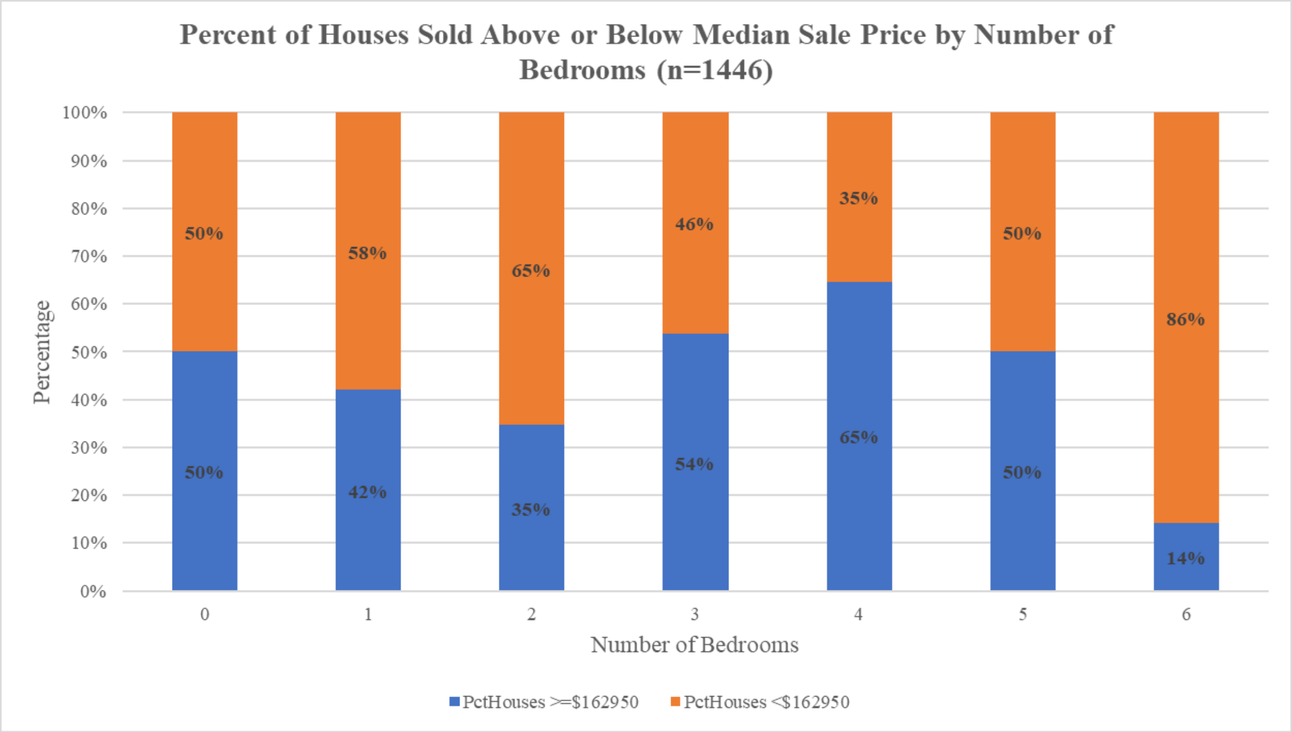
Back to the 100% stacked column chart we go – this chart gives us a lot of information. There does not seem to be an impact due to the number of basement full bathrooms. We should note that there is only one house in the dataset that had three basement full bathrooms, so we should be careful when interpreting this. Including the number of houses for each category is imperative to ensuring that this is analyzed in context. This is a lesson that would be applied to future analysis, such as the final exam.

NumFullBath	PctHouses >=$162950	PctHouses <$162950	Number of Houses
0	33%	67%	9
1	15%	85%	650
2	78%	22%	756
3	100%	0%	31
Total	-	-	1446



We can see a somewhat positive correlation between the number of full bathrooms and sale price. A house having two or more full bathrooms above ground will typically sell above the median sale price, which is $162,950.

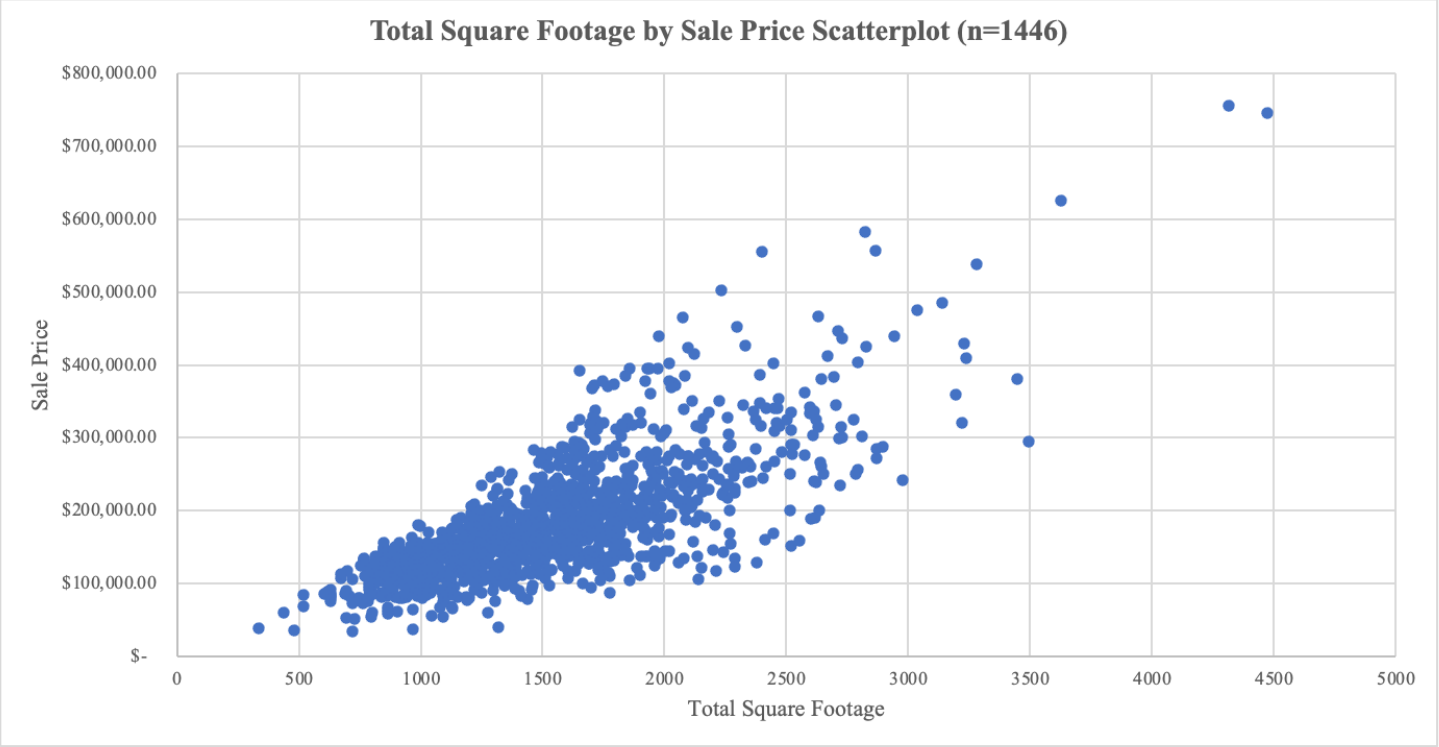
NumBdrm	PctHouses >=$162950	PctHouses <$162950	Number of Houses
0	50%	50%	6
1	42%	58%	50
2	35%	65%	355
3	54%	46%	798
4	65%	35%	212
5	50%	50%	18
6	14%	86%	7
Total	-	-	1446



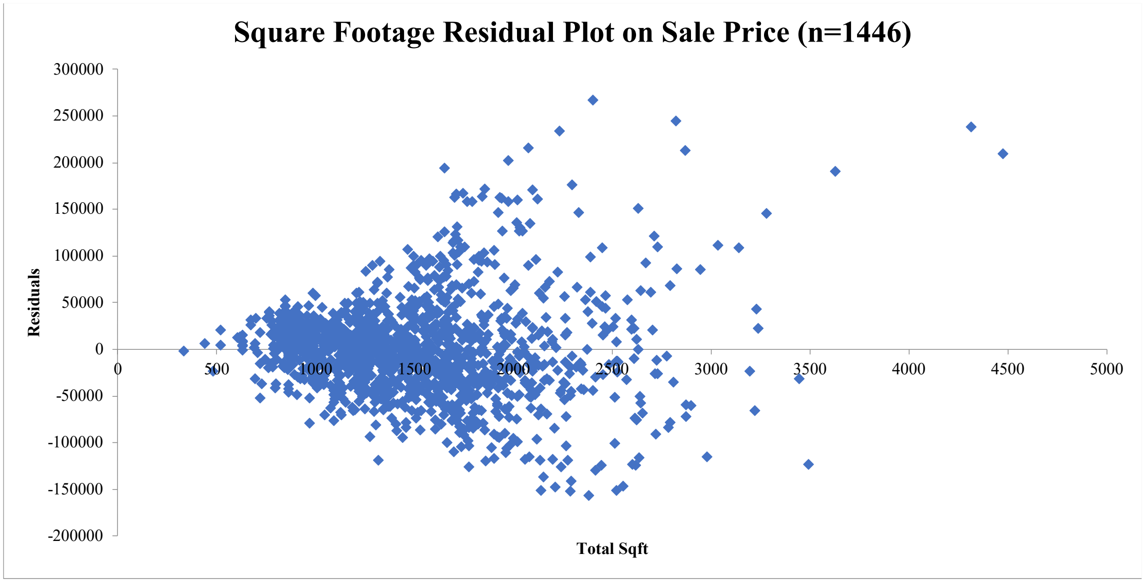
Finally, we have the comparison for the number of bedrooms and the median sale price. This is rather interesting because it seems there is no relationship between the number of bedrooms and the sale price – at least, the sale price does not increase as the number of bedrooms increases. We see that out of the 7 houses that had six bedrooms, only 1 sold above the median sale price, while the rest sold below the median sale price. Again, we need to be careful with this analysis since the number of houses in some of these categories are small. That is why the table is included to ensure transparency.

Bivariate Analysis and Testing

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1639.122294	4267.458944	0.384097964	0.700962391	-6731.960104	10010.20469	-6731.960104	10010.20469
Total Sqft	119.2876668	2.699817466	44.18360438	1.8764E-270	113.9916828	124.5836509	113.9916828	124.5836509



When working with the scatterplot, it is safe to argue that as the total square footage of the house increases, the sale price increases as well. The spread of the scatterplot seems to become less concentrated after the total square footage goes above 2000. Since that is the assumed case, it would be wise to consider seeing if the total square footage is an impactful variable on the sale price. Upon looking at the table above the scatterplot, we see that for every square footage the house adds, the price increases by about $119. Also, there is a t-test for significance. Since the t-statistic is around 44 standard deviations above the mean, the p-value is outlandishly small, and zero is not in the 95% confidence interval, then the total square footage is a significant variable to predict the sale price. Consider the residual plot below:



In the above residual plot, there was a noticeable shift in the accuracy of the model as the total square footage increases. We have found that the total square footage is a major factor in predicting the sale price, but it is worth noting that when the square footage is about 2000 or more, some of the residuals are large. This tells us that the predicted sale price of these houses has a large error – to be specific, the R^2 value is 0.57. So, the change in square footage can be modeled for about 57% of the data. This is important to keep in mind when doing this analysis; however, it is expected since there are more variables to consider which would improve the model.

Multivariate Analysis and Testing

Table 2

Variable	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Total Sqft	84.04630615	2.788790002	30.1371943	4.5505E-155	78.57577371	89.5168386
OverallQual	23907.05617	952.9192552	25.08822866	1.6446E-115	22037.79542	25776.31692
Intercept	-164076.4439	7356.310431	-22.30417618	7.00324E-95	-178506.6932	-149646.1946
Yrs Since Remodel	-353.1007444	53.09687689	-6.650122664	4.14825E-11	-457.2563773	-248.9451114
Total Baths	9691.017773	1567.855774	6.181064567	8.27689E-10	6615.488286	12766.54726
DCulDSac	16618.1442	3353.66955	4.95521218	8.0812E-07	10039.53553	23196.75287
DCorner	1276.000691	2220.709304	0.574591501	0.565657477	-3080.176106	5632.177489

The above table is a regression test for each variable that predicts the sale price. Notice how the number of variables decreased from the quantitative and qualitative analyses earlier in the paper. This is because we have condensed the data to use only useful predictors and make the analysis much more intuitive. In table 2, the data is sorted in ascending p-value, which tells us which variable had the most impact on predicting the sale price. It is worth noting that the total square footage had the most impact on predicting the sale price – the only variable in this table that did not have a significant impact was if the house was a corner lot, which we will discuss later. Since we have discussed the total square footage in the previous section, and the outcome is similar in this model, we will focus on the other variables and then form a model to predict sale price. Note that the intercept is the initial value, so there is no need to deeply analyze it in this case. For this regression model, the R^2 value is 0.82. So, the change in the sale price can be modeled for about 82% of the data.

First, we will consider the overall quality of the house. From the data table generated from the regression test above, the overall quality is a significant variable in predicting the sale price of a house. This is because the t-statistic is 25 standard deviations above the mean, the p-value is much less than 0.05, and 0 is not in the 95% confidence interval. To be more specific with the t-statistic, the hypothesis test was to see if the value of the coefficient *could* be a nonzero value. Since the test resulted in a value in the rejection region, which is 25 standard deviations above the mean, then we concluded that the coefficient value is nonzero. So, for every increment in overall quality of a house, the value of the house increases by about $24,000.

The next variable is the number of years since remodeled. Since the t-statistic is in the rejection region, the p-value is much less than 0.05, and the 95% confidence interval does not contain 0, then the number of years since remodeled is a significant variable in predicting the sale price of a house. Note that the coefficient generated is a negative value. So, for every increment in the number of years since remodeling, the value of the house decreases by about $350. This makes sense since if the house has not been remodeled for a long time, the house is probably not in the most pristine shape, which decreases the value slightly.

After the number of years since remodeled, consider the total number of bathrooms. In this case, the t-statistic is more than six standard deviations above the mean, implying it is in the rejection region. Also, the p-value is much less than 0.05, and the 95% confidence interval does not contain zero. Since that is the case, the total number of bathrooms is a statistically significant value in determining the sale price of a house. In fact, for every bathroom the house has, the value of the house goes up by almost $10,000.

Next, consider the 'DCulDSac' variable. This is a binary variable which identifies if a given house is in a cul-de-sac. When conducting the regression test for this variable, we have found that the t-statistic is almost five standard deviations above the mean, which implies it lies in the rejection region. With that said, the p-value is significantly smaller than 0.05 and the 95% confidence interval does not contain zero. So, this variable is also significant in determining the price of a house. The coefficient calculated from this regression analysis is $16,618 – implying that the value of a house increases by that amount if it is in a cul-de-sac.

Finally, consider the 'DCorner' variable. Like the above variable, this variable identifies if a given house is a corner lot. The regression analysis yielded that this variable is not very significant in determining the price of a house. The value of the t-statistic is only 0.57 standard deviations to the right, which certainly does not lie in the rejection region. Further, the p-value is just above 0.05, and the 95% confidence interval does contain zero in the interval. Since that is the case, the significance of this variable is not strong in predicting the price of a house. With that said, the value of a house increases by about only $1,300 if it sits on a corner lot, which is not a major change in the house's value.

In the above table, the most significant variables to determine the sale price of a house are the total square footage, overall quality, and number of years since remodeled. The above table is listed in order from smallest p-value to greatest p-value. So, these are the most significant. It would be interesting to use the most significant variables within this model to make another model. With that said, we will conduct another regression test while only using these variables to see how the model turns out.

Table 3

Variable	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Total Sqft	68.9134636	2.514514572	27.40626933	1.8399E-133	63.98096549	73.84596171
OverallQual	27127.40503	1052.773964	25.7675493	9.5021E-121	25062.2726	29192.53746
Intercept	-77864.46641	6276.476292	-12.40576126	1.18516E-33	-90176.46801	-65552.46481
Yrs Since Remodel	-446.5320178	58.08383819	-7.687715408	2.75518E-14	-560.4698829	-332.5941528

Like the larger table in the previous regression analysis, we have the variable with its t-statistic, p-value, and 95% confidence interval. When performing another regression analysis with just these three variables, we can see that the number of years since remodeled and overall quality of a house carries more weight. The total square footage does not carry as much weight as it did in the previous analysis. For further information, the R^2 value of this regression is 0.77, implying that the change in the sale price can be modeled for about 77% of the data – this is a bit weaker than the previous model, but we will discuss the significance after this analysis.

Again, the order of this table is sorted from smallest p-value to largest p-value. Starting with the total square footage, we have found that this variable is significant for this model. The t-statistic is 27 standard deviations to the right, implying it lies in the rejection region. Also, the p-value is significantly smaller than 0.05 and within the 95% confidence interval, zero is not contained within the interval. Since all these conditions are met, the total square footage is significant in determining the sale price. For every square foot the house has, the value of the house goes up by $69.

The overall quality of the house was also found to be significant in this regression test. The t-statistic came out to be about 26 standard deviations to the right, which lies in the rejection region. Also, the p-value is significantly smaller than 0.05 and within the 95% confidence interval, zero is not contained. So, for every point gained for overall quality of a house, the value will increase dramatically – about $27,000.

For the final variable, number of years since remodeled, it was also found to be significant. The value of the t-statistic lies 7.7 standard deviations below the mean, which is in the rejection region. With that, the p-value is much less than 0.05 and zero is not contained within the 95% confidence interval. In this case, the greater the number of years since remodeled, the lower the price of the house gets. This is like our previous model and, again, makes much sense. For every year after the house has been remodeled, the price of the house decreases by about $450.

Summary of Findings

By now, we have conducted three regression tests and developed three models to best model the price of a house. The first model was just using the total square footage of a house. While we only used one variable for this model, it was a surprisingly decent model with an R^2 value of 0.57. So, the change in the sale price could be modeled for about 57% of the data. This is not that bad; however, there are many more variables that influence the price of a house. Making these variables work together could be key in making a model that is accurate and not overbearing. After the first model, we moved to the second model which used all the significant variables in the dataset to provide a more accurate model. The attempt was successful – the model resulted in an R^2 value of 0.82, which is a 5% increase in value with respect to the number of variables added. This is outstanding – most of our data can be modeled with this; however, one thing we must consider is the number of variables. There are a total of six variables within this model, which can be a bit to manage. We should keep this in mind when considering the final model.

The final model was an interesting one – the model was derived from the most significant variables from the second model. Within this model, the R^2 was found to be 0.77. With respect to the number of variables present, this is a 1.5% decrease in value compared to the second model. This is a bit less than the second model, but we should consider the opportunity cost of these models. The first model is not bad, but it would be wise to see how more variables impact the sale price since we have the data at our disposal. The real debate is between the second and the third model – which one would be better to use? We should note that the second model has more variables than the third model – the third model has half the variables that the second model does. The limited number of relevant variables makes the model more manageable to use, which is a huge bonus. On top of that, as stated earlier, using the third model would be about 1.5% less valuable than the second model; however, the opportunity cost of that is worth it. We need to remember that we are trying to maximize profit for this real estate company, so minimizing the number of variables is critical in maximizing profit. Therefore, the third model would be the most reliable and reasonable one to use. To be explicit, the company may use this formula, which took twenty-two pages to develop:

$$SalePrice = -77864.47 + 68.91(Total\ SqFt) + 27{,}127.41(Overall\ Quality) - 446.53(Years\ Since\ Remodeled)$$