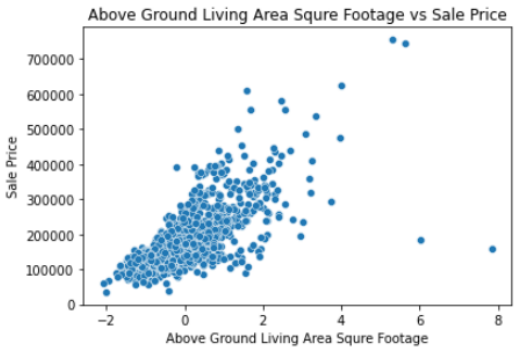
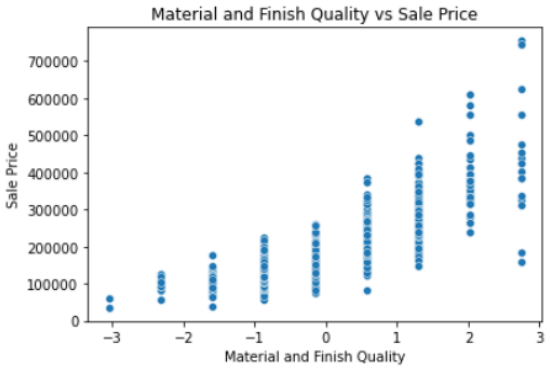
Introduction

I am using a dataset that has a bunch of information about houses such as their condition, location, and features. I want to predict how much money the houses are getting sold for using linear regression. Linear regression uses the "least squares" method to find a "line of best fit" for a dataset. This line is calculated by finding the distance from each point to the line, and whichever line is generally closest to the data as a whole is the line of best fit. Using the variation in the data compared to the line, an R-Squared value is found. The R-squared value tells us how much of the variation in the dependent variable is explained by the independent variable. To find out if the model is useful, we find a p-value for the R-squared value. Depending on the p-value and critical value set before the experiment, we can determine whether or not the model is useful.

Because I am using a linear regression model, my first step in preprocessing was to cut the dataframe down to only the numeric data, as any categorical variables wouldn't work in a linear model. The next preprocessing step was to see where any null values were, and the best way to deal with them. Because there were only a few hundred and the dataset still contained well over 1000 samples without null values, I dropped any sample that contained a null value. Because I am using linear regression, I used the standard scaler to put everything on the same scale, meaning that no variable would be more meaningful than any other simply because it had a large value.

Experiment 1

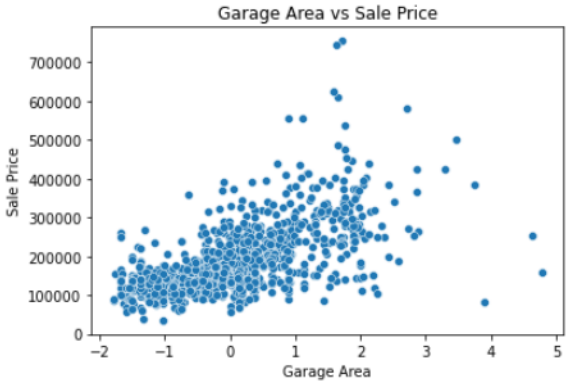




The first experiment conducted was a linear regression model based on every column that contained numerical data, both categorical and continuous. The two most important variables from the first model were "Overall material and finish quality" and "square footage of the above ground living area." The visualizations with each compared to Sale Price are shown above. While both are numerical variables, it is easy to see which variable is categorical and which one is continuous. In our OLS regression results, we see that both these models are statistically significant, with p-values of ~0, and the root mean square error(RMSE) is 38228.48.

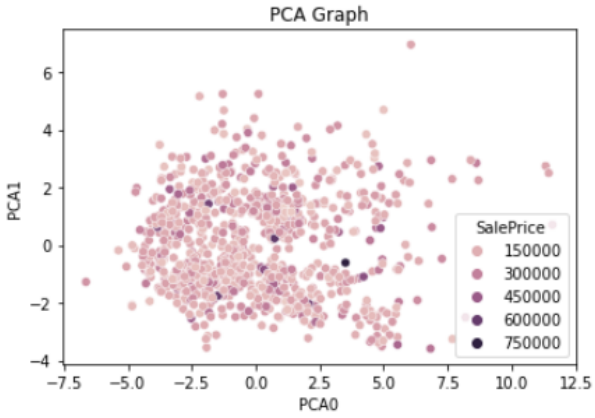
Experiment 2

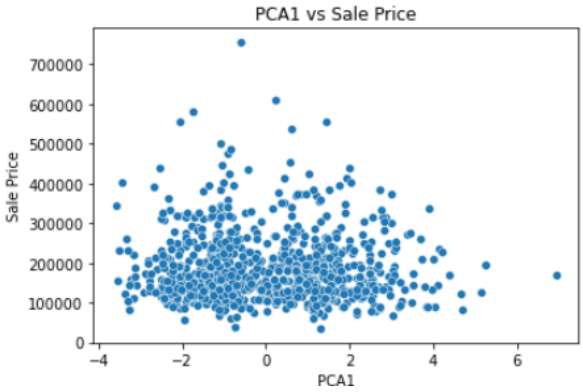
In the second experiment, I want to solely use continuous values. In order to do this, I dropped the most important categorical variables until the two most important variables were both continuous. Above ground area is now the most important variable, and garage area is the second most important. The visualization showing Garage area compared to Sale Price is shown to the right while the graph for Above ground area is shown in experiment 1. Just like the previous experiment, both variables had p-values of ~0, and this model had a RMSE of 57845.26. Because there were multiple, more important, features not being used, a higher RMSE is expected.

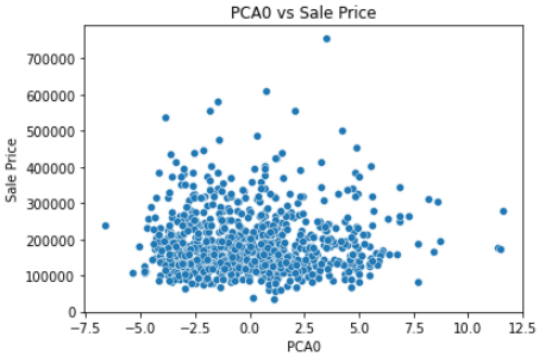


Experiment 3

The third experiment was done using Principal Component Analysis(PCA), as opposed to any specific type of variable. PCA reduces the dimensions of a dataset to a desired number of dimensions by combining variables as much as needed. I combined all 38 numerical variables into 2 dimensions, and created three visualizations. The first visualization, shown to the right shows PCA0 and PCA1, with a hue of the Sale Price variable. There isn't much correlation between the two dimensions. The visualization shown directly below the PCA Graph is between the first PCA dimension, "PCA0", and Sale Price. The visualization to the left is between "PCA1", the second PCA dimension, and Sale Price. What these three graphs tell us is that it is hard to combine 38 dimensions into two. Because of the "black box" nature of PCA, I was unable to figure out how to find the same metrics to measure the model. However, we can clearly see that the models aren't very accurate when compared to the more important variables alone.







Impact and Conclusion

The findings of this project are important for current homeowners looking to possibly renovate and sell their homes, or looking for ways to improve their home currently that will contribute to it appreciating in value. I'm not sure that this will only affect single homeowners though. Big corporations that buy up large amounts of the housing market could use this data so they know what to focus on when building homes to either rent or sell for profit. A couple ways to improve this project would be to find ways to quantify more variables, which would allow us to see a bigger picture of what makes a home valuable. Another potential improvement would be an increase in number of PCA dimensions, and potentially finding an amount that turns up more correlation between Sale Price and a given dimension or dimensions.