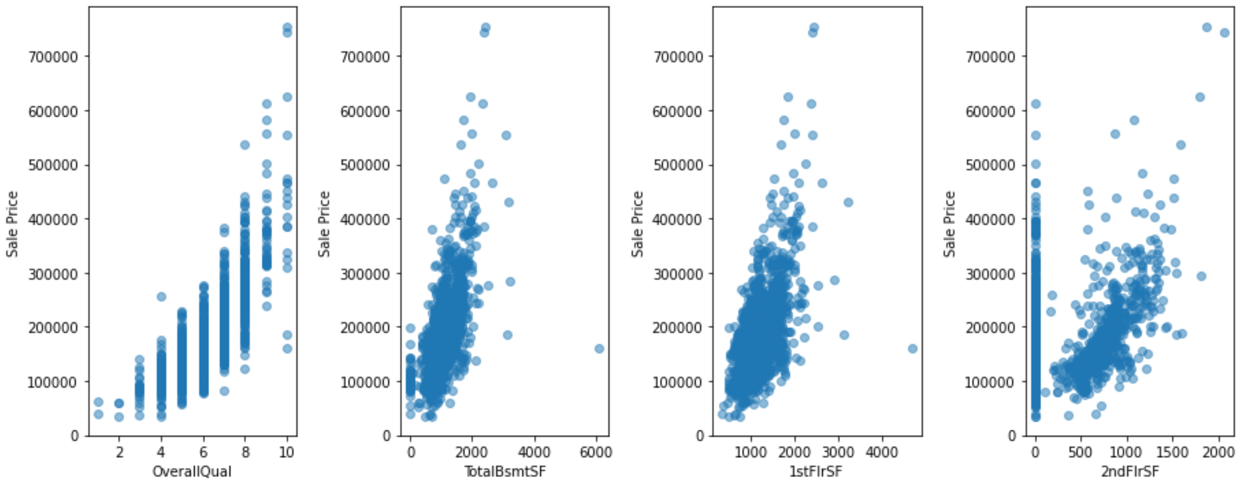
# Predicting Housing Prices

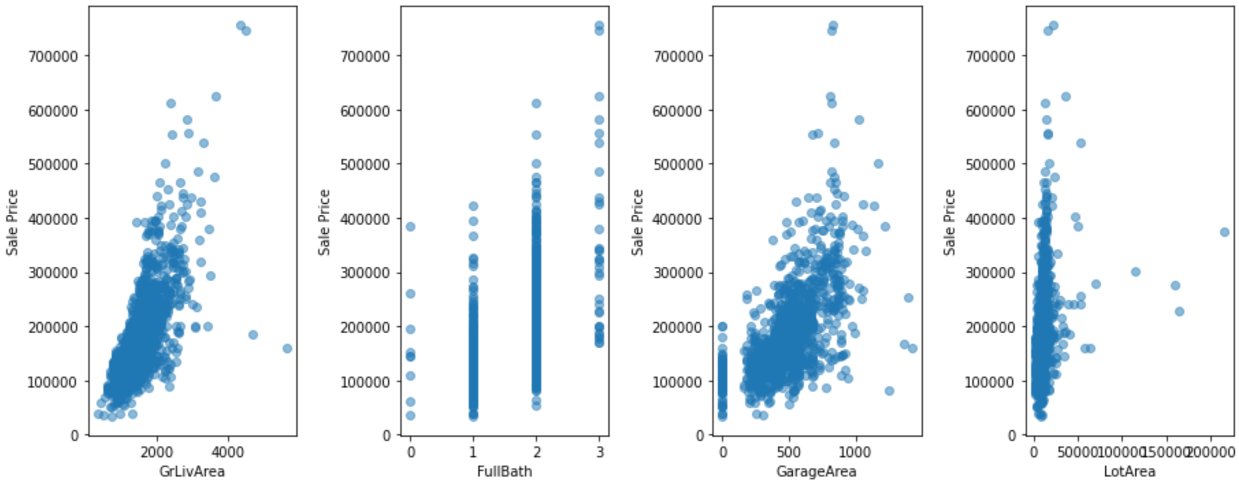
One of the most important things when looking at selling or buying a house is to have an appropriate price. If you are trying to sell a house, this can be critical. Overpricing a house can result in a lack of offers. You can eventually lower the price to attract offers, but in the meantime you incur carrying costs like taxes, mortgage, utilities, and insurance. Underpricing a house will sell the house faster, but you won't get full value for the house. For someone buying a house, you want to know if the price you see for a house is a good price, or at least really fair. You can look at comparables in the area, but there could be only a limited number of those which can make trying to get a good idea of the price difficult. What would be helpful is a way to take features of the house and use that to generate a price. This would be especially helpful if there aren't really nearby houses comparable to the one being sold.

To be able to do this, we can start by taking a relatively large set of houses sold, with several features noted as well as their price. Then we can use machine learning to build a regression model. This would allow us to put in the features of a given house and use this model to predict the price of the house.

I tok data from **https://www.kaggle.com/c/house-prices-advanced-regression-techniques/** which had a set of data that included prices and another set that did not. I used the one that included prices to build a model. I was then able to apply that model to the other houses. Once those other houses sold, we could see how well that model worked on that set.

Before trying to actually create a model, it was useful to look through the data to see if any trends stood out. One of the first things that was examined was to see which (if any) features appeared to strongly correlate to the price of the house. Not surprisingly some of the features that had a relatively strong correlation to price were the overall quality of the house, the total living area, and other sizes. We can see how some of these appear to correlate in the following figure:

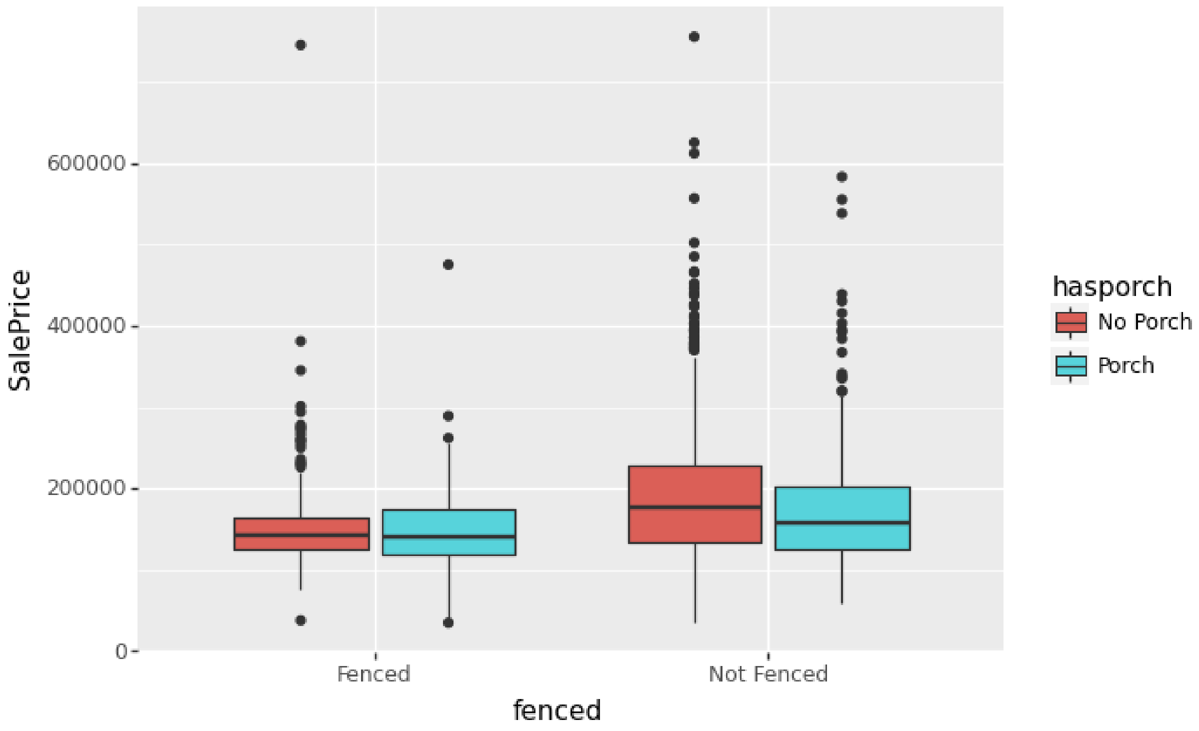




We can see that as the rating of quality of the house goes up, so does the price. Square footage of the basement, first floor, second floor and over all living area definitely contribute. The strange looking vertical line for 2nd floor footage represents the single floor houses.

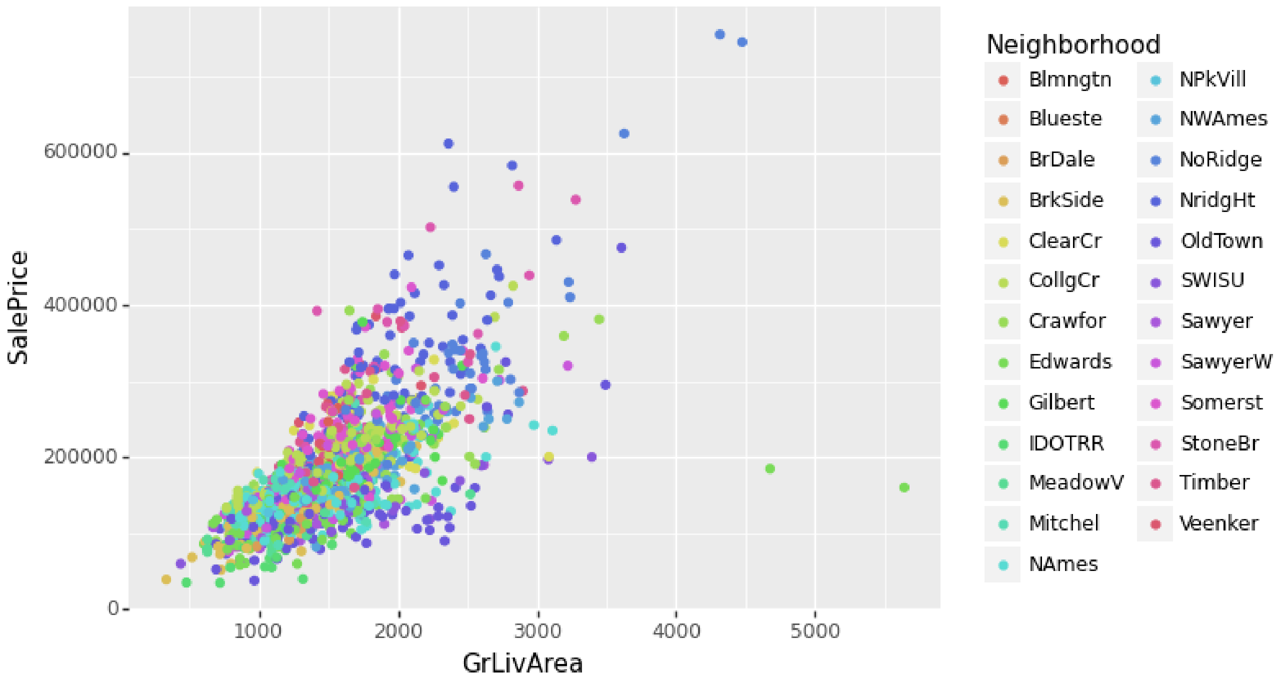
With things like lot area and living area so strongly correlated overall, we used these to explore to see if some other features affected the slope in those instances. In other words, we wanted to see if the price per square foot looked like it varied based on those features. For many features, there wasn't any obvious pattern, but we found a few that were easy to see.

One of the first things we looked at was whether having a fenced in yard or a porch seemed to affect the price of the house. To do this we made a box plot of the prices separating by having a fence or not and having a porch or not. The following figure is what we got.

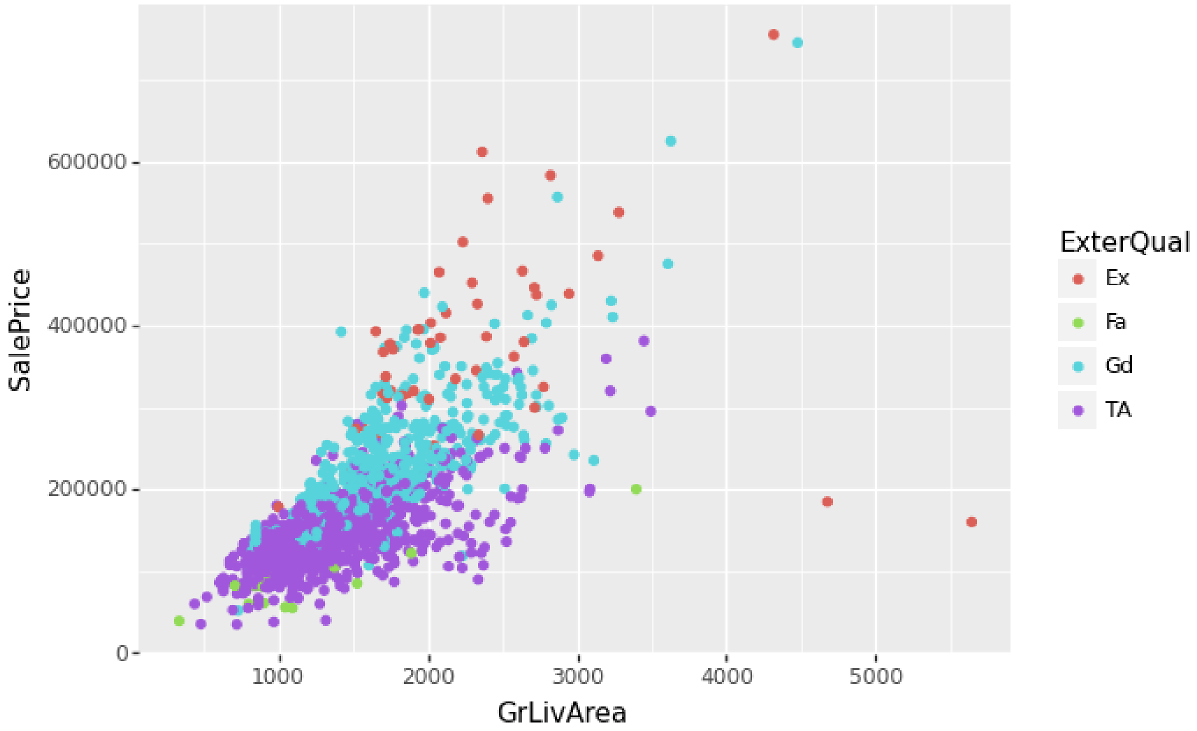


Surprisingly houses with fenced in yard didn’t appear to have prices much higher than the ones without. Part of the issue could well be that there weren’t many houses with a fence so it was too small a sample to be meaningful, but we couldn’t really see any real influence of having a fenced in yard.

We also wanted to see how location affected the price as it related to area. There were some difference, but with many neighborhoods and not many houses in each it was hard to see anything particularly standout as we see here:



There are a few that seem to be steeper slops but there was a lot of overlap. One feature that we could easily see visually affect this graph was the quality of the house.



Here “Ex” is excellent quality, “Gd” is good quality, “Fa” is fair quality, and “TA” is “typical/average quality”. We can easily see that excellent or good external quality homes sold for more per square foot than the fair or typical/average external quality homes. There were other features that had a similar effect, but generally those correlated to the external quality.

For our first attempt to build a model, simple linear regression was used. To apply it, we didn't include any features where the vast majority of the homes were in one class. For example, all but 26 of the room materials were listed as Composite Shingle. There was just not enough useful data to make use of this category. The exterior quality was definitely going to be an issue. Two approaches were tried for this. The first was to split the homes into two groups, one consisting of the Good and Excellent exterior quality homes, the other with the fair and typical/average. The other was to convert that feature to numeric where 0 was fair, 1 was typical/average, 2 was good, 3 was excellent.

For the first approach we just treated each separately. However , after trying to tune how many features to use, we were unable to get a $R^2$ score for our test data above about about 0.69
In our second approach where we used the numeric conversion, we did not do much better.

After this we looked at the lightgbm model. This is a gradient boosted model based on decision trees. We didn't drop any categories, and since this model can use categorical features, we could just leave all our features as it is. After a little preprocessing to get the data set to use, we were able to achieve an $R^2$ score of about 0.88 on the validation data. This was a great improvement and was the model that we used in the end.

With this model, a realtor could now take a client's home, input this data and get a starting point for price. There are always factors unrelated to the house itself that might be relevant, like the urgency of selling the house, but at the very least the realtor and seller would have an informed starting point. Also, if someone was looking for homes, they could take a home for sale, and use this model to get an idea if the listing price was appropriate. Again, external factors might affect this , but for both a seller and a buyer, having a good idea to start with would help.

There are of course limitations to this specific model. For one thing, it's definitely limited to a specific area. There are variations by neighborhood, but prices in a major metropolitan area will be far different than those in a rural area, and will be affected by different factors. One approach might be a nationwide database where houses around the country are input. For something like that probably demographic data would be useful or maybe zip codes would be useful. A local realtor in a different area could collect the same type of data for their area and then run the same type of model there. Another future improvement would be more precise information about when the house was sold. This data only had the year the house was sold, but the markets change over the course of a year, so having that might allow better estimation. Anyone who really wanted to use this over time would be well served to rerun the model every so often with new data to really help capture what is happening currently. This issue is not unique to this approach. In areas with relatively small markets it's quite possible that the comparable houses sold would not be that recent, and the fluctuations of the real estate market over time wouldn't be captured there either. This approach should be a great asset to anyone buying or selling a house.