Capstone 1 - Data Story

Home Sale Prediction - Ames, Iowa

Initial Assumptions About The Data

Now that we've cleaned the dataset (see Data Wrangling pdf), its time to evaluate some univariate and bivariate relationships using exploratory data analysis. The target variable for this project is home sale price for homes in Ames, Iowa.

Home sale prices can be affected by a number of variables. It's clear from the analysis that follows that many if not all of the variables in this dataset have some influence on sale price. There are however a few initial assumptions we can make based on some basic knowledge of real estate.

When it comes to real estate there is the old saying 'location location location'. Location for this dataset is established by neighborhood. Let's make the initial assumption that neighborhoods will have a significant influence on sale price. Some will have better schools and less crime and this will directly affect the sales price of homes.

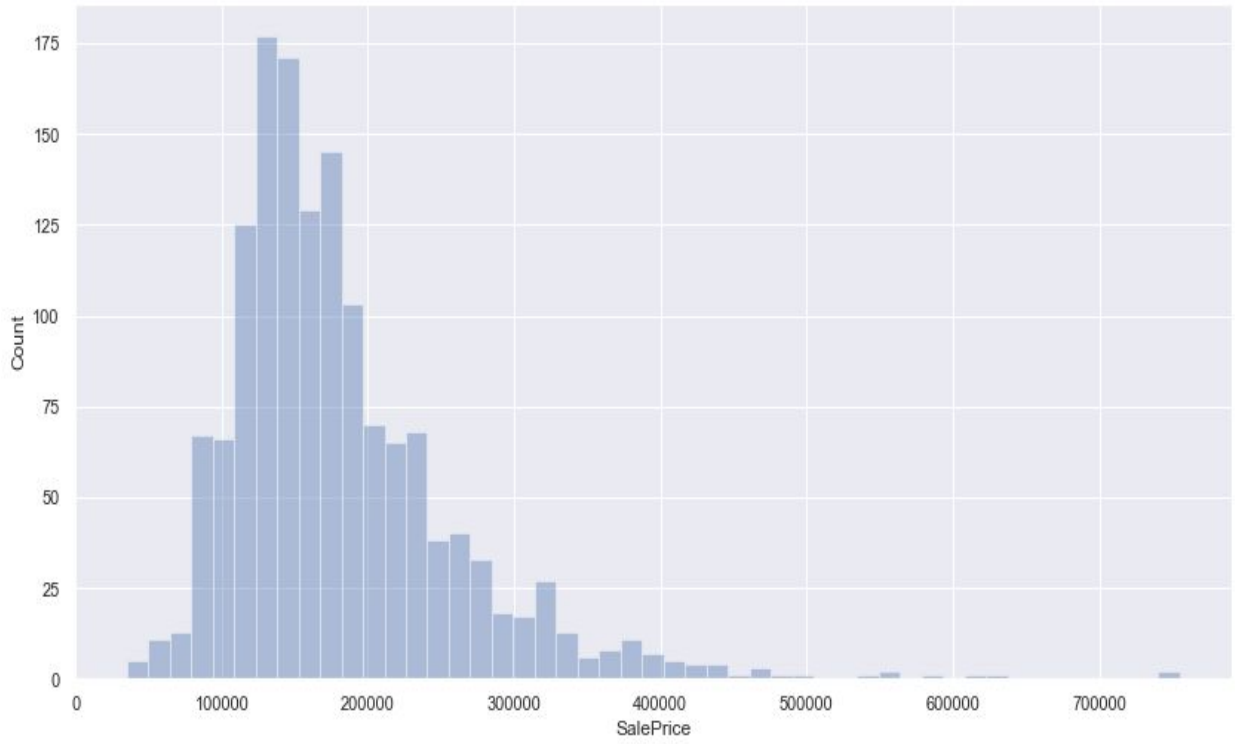
The other variable of interest here will be the square footage or overall size of the home. In this dataset there are many variables which describe the home size. These include: Lot area, Livable square footage, Basement square footage, Garage area in square feet and several others. My initial instinct is that size of the home will be the most influential feature in this study.

One final feature to examine would be the date (month, year) in which a home was sold. The goal here is to examine if there is any pattern in volume of home sales which could potentially affect sale price.

Lets begin to explore the data and test some of these initial hypotheses.

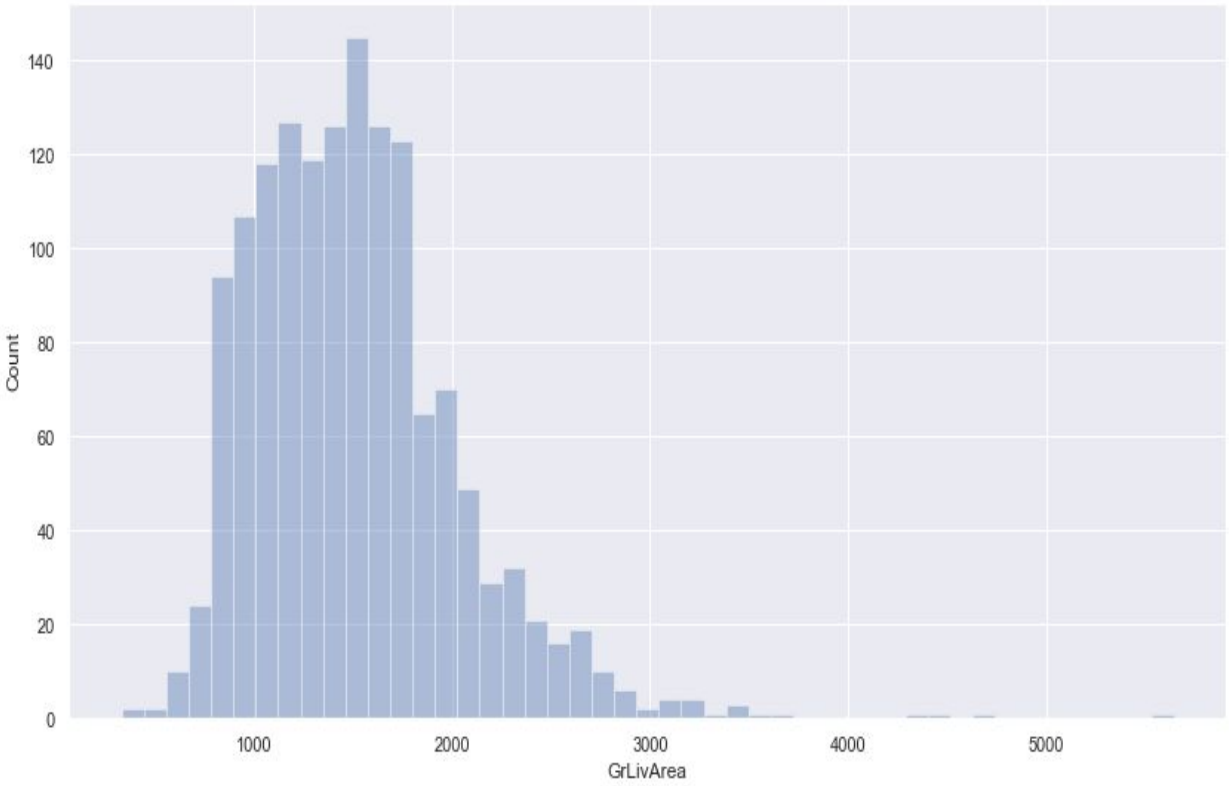
Preliminary EDA

Let's take a look at the target variable and see what its distribution looks like:



The distribution seems skewed to the right, with what look like several outliers. Is this data valid? Could there be a mistake or were these homes truly valued so much higher than the rest of the homes in this dataset?

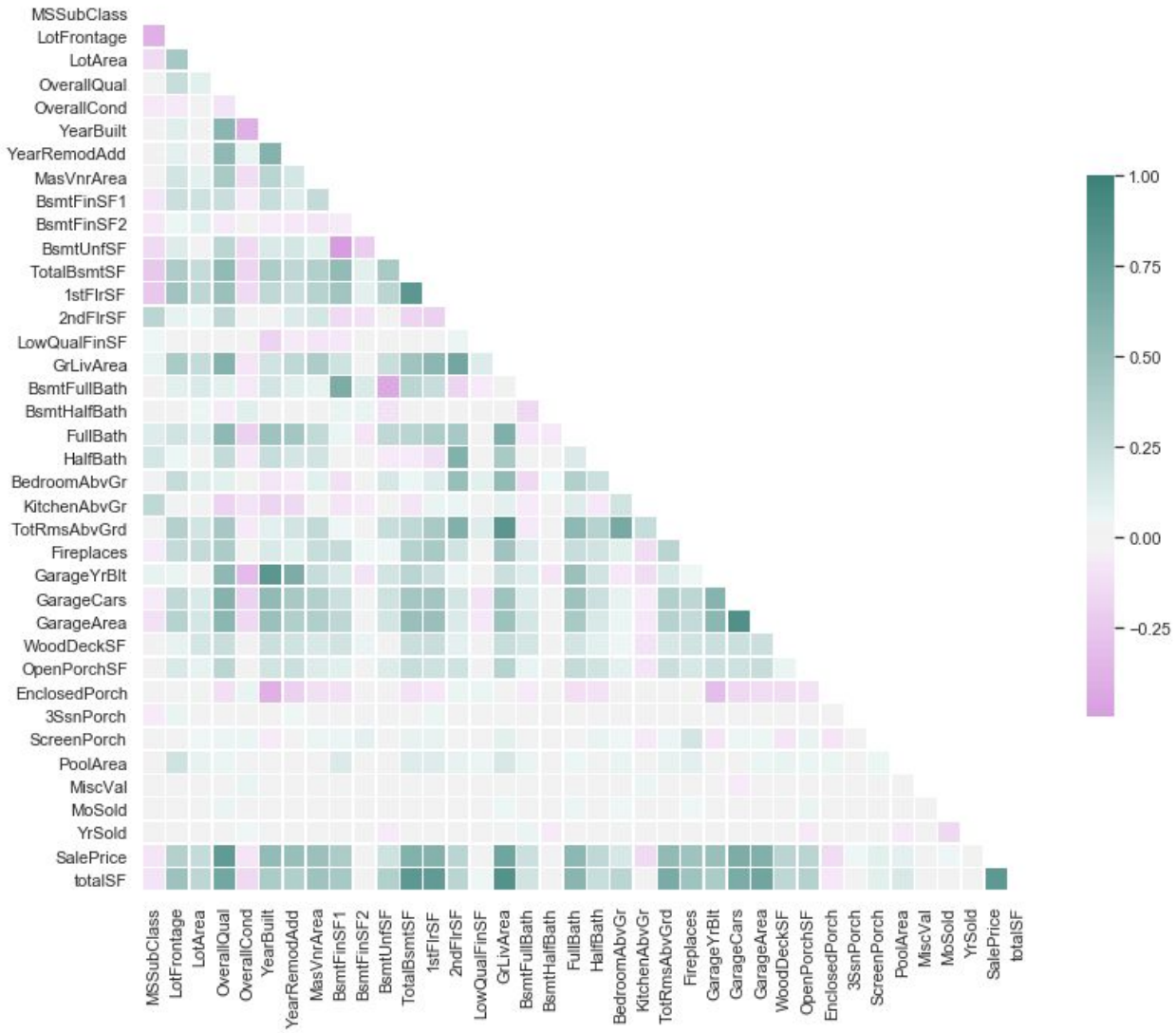
Let's take a look at what seems to be the most obvious influencers on sales price, livable square footage (the variable GrLivArea). The goal here is to see if we see a similar distribution to that of sales price. If so, this would indicate perhaps that this skewed sales price distribution might in fact be representative.



This is interesting, it looks like the living area (or square footage) also has some outliers and is skewed to the right just like sales price. So these 'outliers' might actually be validated by significantly larger square footage.

Numerical Correlations to Sale Price

Let’s dig deeper into the numerical data types by evaluating their correlation with respect to sale price:



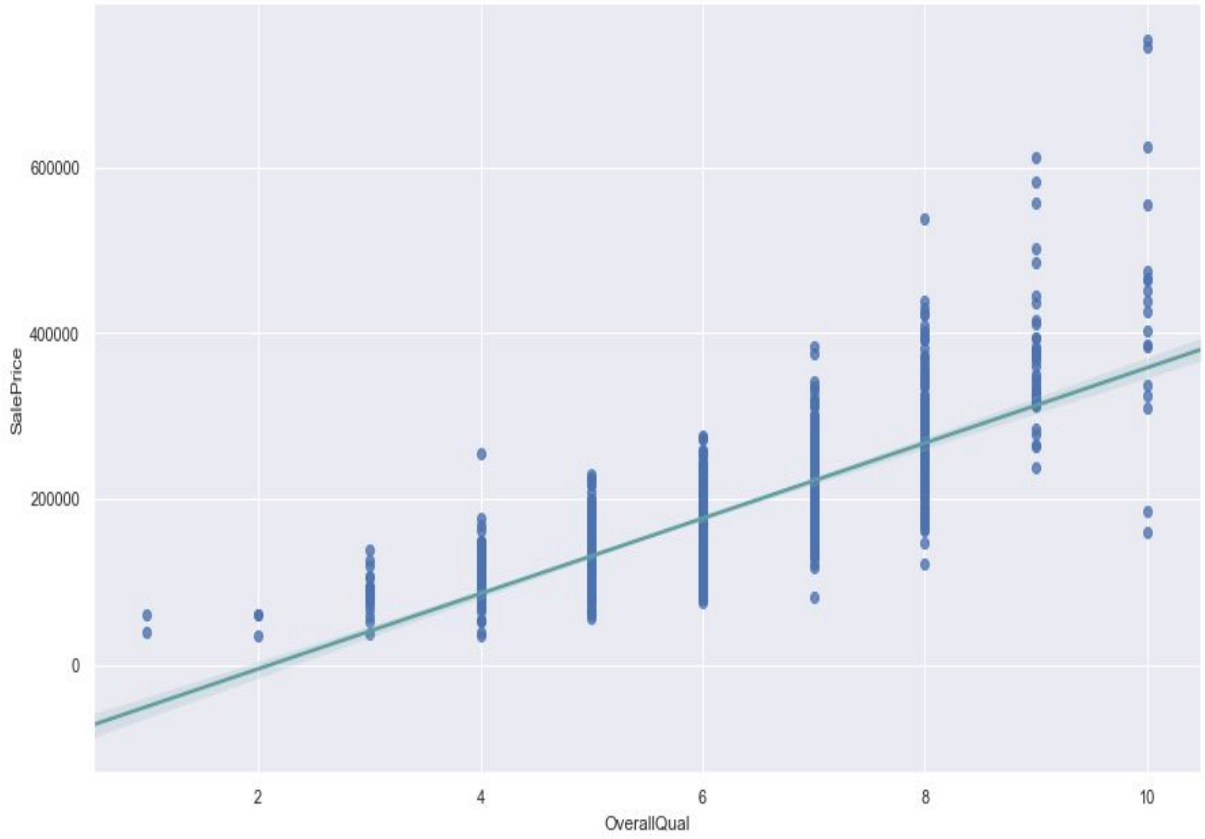
The parameter with highest correlation to sale price is in fact ‘OverallQual’. However we do see that right behind this feature are ‘GrLivArea’ (livable space in square feet), ‘TotalBsmtSF’

(square footage of the basement) and ‘GarageArea’. So this seems to support the initial hypothesis that square footage has the largest impact on sale price.

Examining Numerical Features with High Correlation to Sale Price

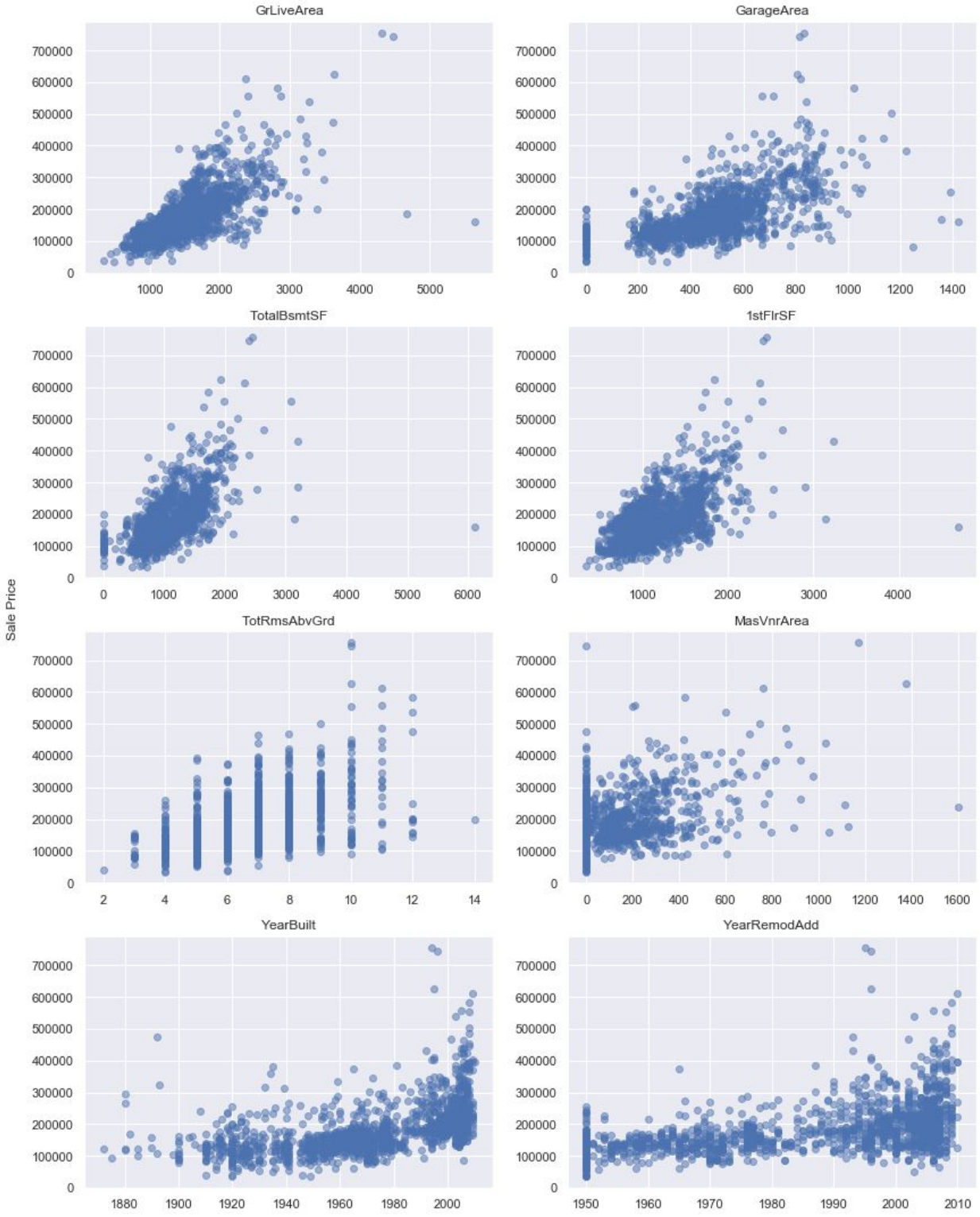
Overall Quality of Homes

Let’s take a look at the overall quality feature represented by the ‘OverallQual’ column. This feature has the highest correlation to sale price so we expect to see a nice trend here:



For any given quality score (ranging from 1-10) we see a fairly wide range of sale prices. However there is a clear trend here. Through the means of this correlation matrix we have discovered an unexpected feature which seems to have a fairly linear relationship with sale price.

Let’s now take a look at some of the other features who had high correlation to sale price in hopes of discovering more about this dataset. Below is a plot of several numerical features vs. sale price:

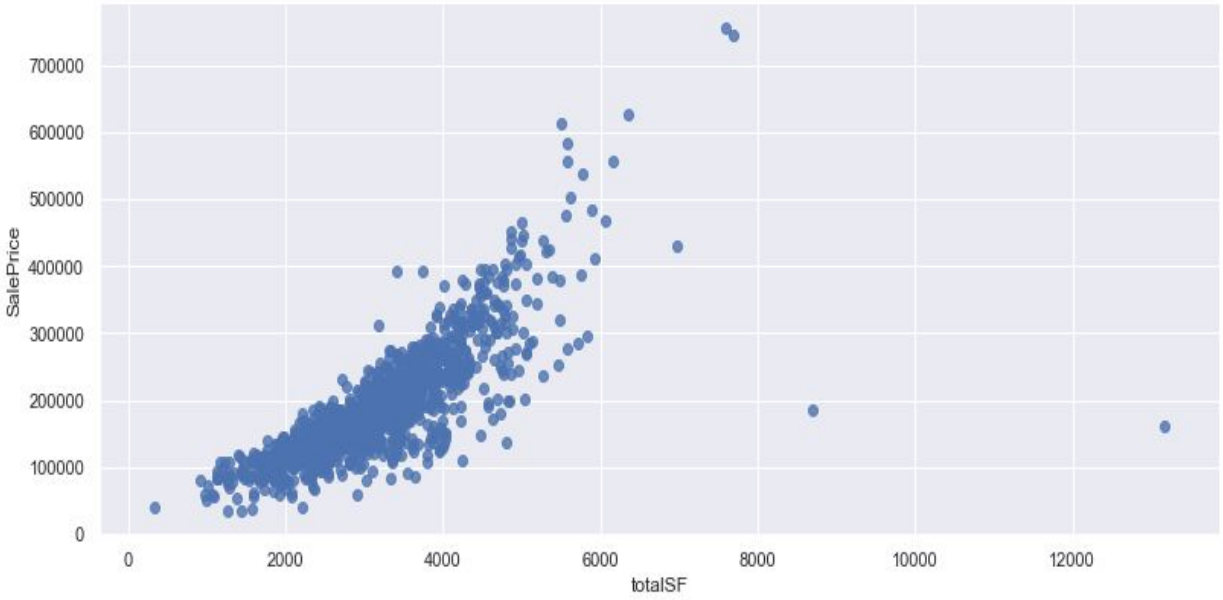


Total Square Footage of Homes

From this result it's clear that the parameters describing usable square footage of the home have a strong correlation to sale price. Let's do a bit of feature engineering to see if we can gain a better understanding of how strongly these features are correlated to sale price
Let's define a new parameter 'TotalSF' which we can set equal to the sum of GrLivArea, TotalBsmtSF and GarageArea. These three parameters essentially describe the total usable square footage of a home. Let's see what the relationship to sales price looks like when we combine them as follows:

totalSF = GrLivArea + TotalBsmtSF + GarageArea

When plotted against sale price, we can produce the following visualization:



This seems to be a very positive correlation. From first inspection there appears to be an almost exponential relationship between sale price and totalSF. This really validates the hypothesis and my initial assumptions about the dataset.

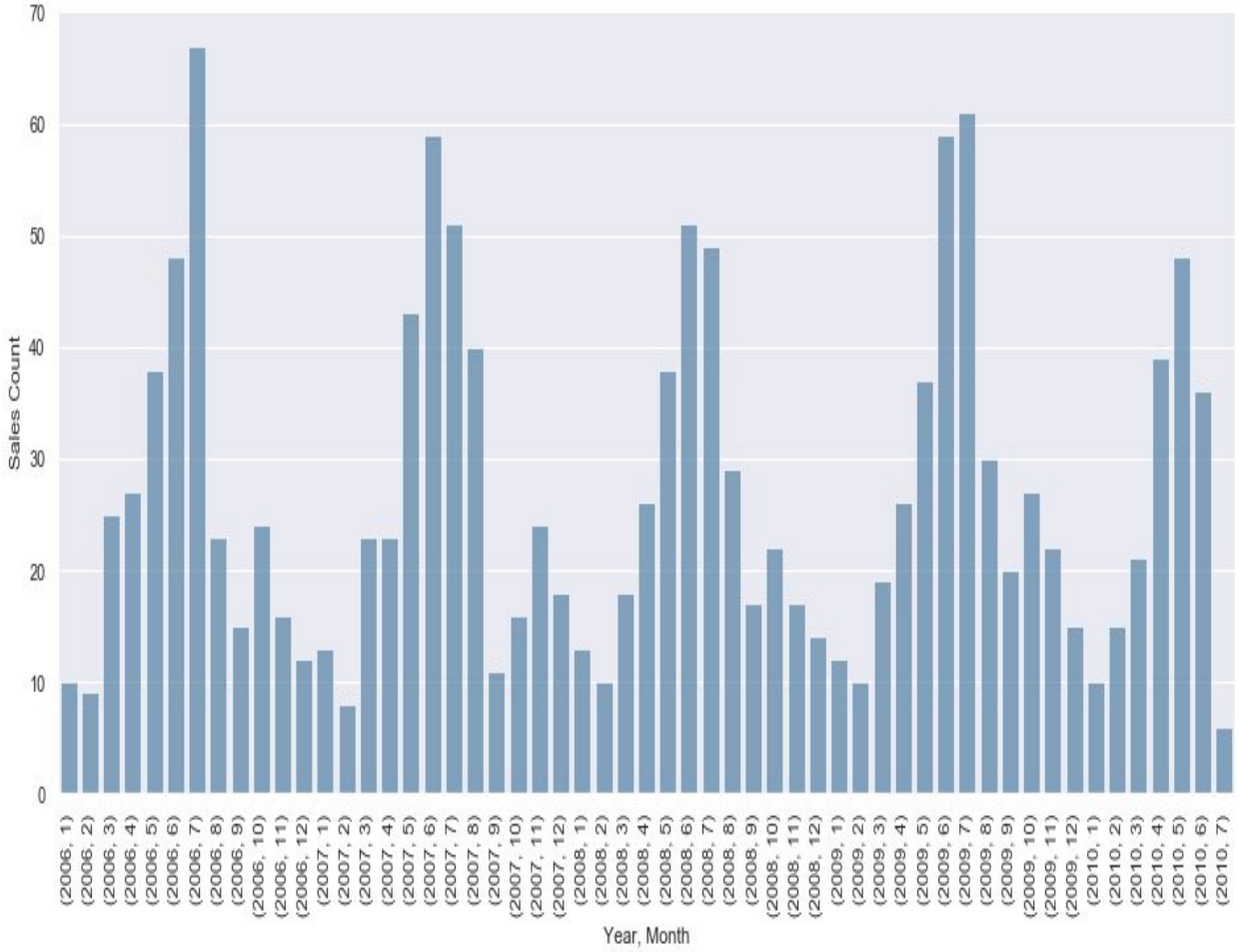
What's even more interesting here is that if we re evaluate the correlation coefficients with the totalSF parameter in the datasets, we obtain the following result:

totalSF(.8) > OverallQual(.79)

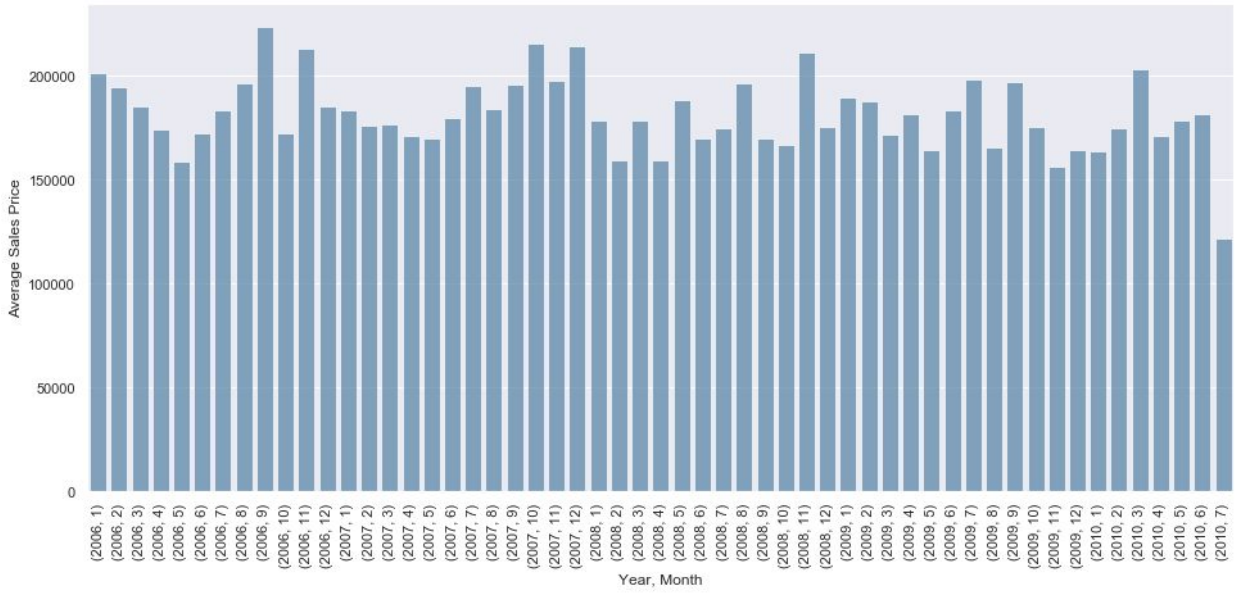
So it can be stated that the feature of highest correlation in this dataset is the livable area or square footage of the home.

Evaluating Sales Volume

Sales volume can be defined as the number of homes sold in a given period. In this case we can combine the month sold (‘MoSold’) and year sold (‘YrSold’) parameters and then aggregate data by month sold, year sold to establish a count of homes sold during that period. This produces the following result:



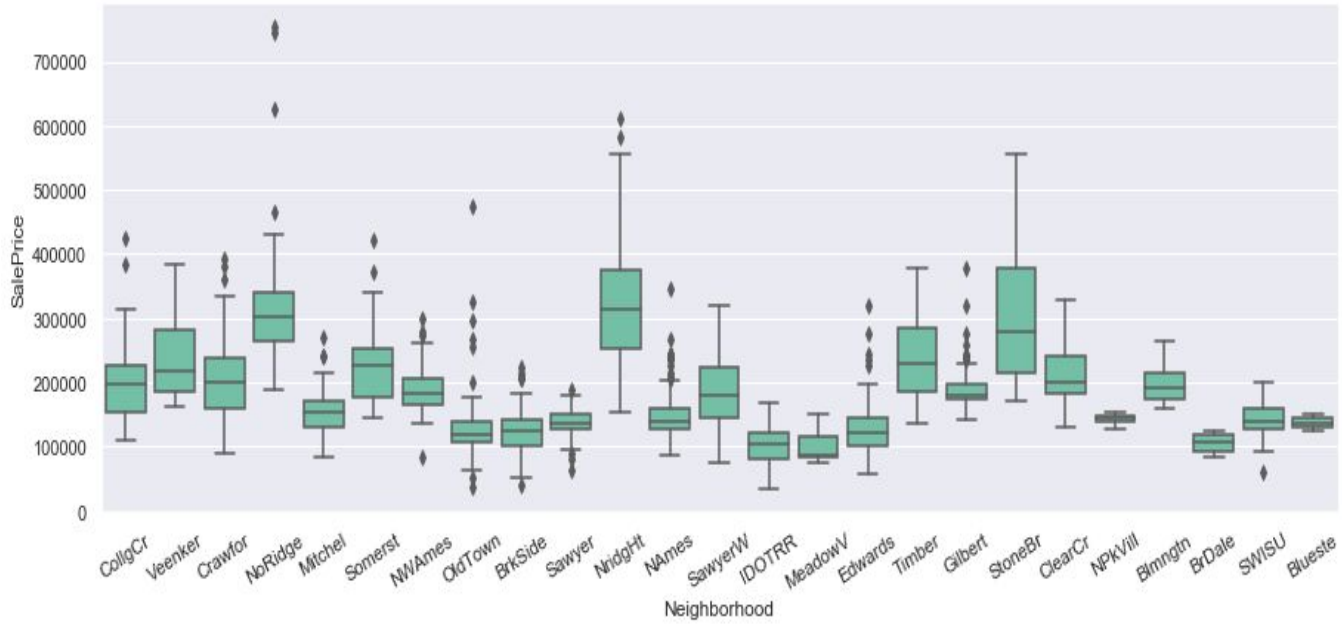
This is quite a fascinating visual. It’s clear that there is a very cyclical pattern to sales volume with peaks every summer and troughs in the winter months. Using this new feature let’s see if this same cyclical relationship is mirrored on sale price:



From this result there seems to be a trend, but not nearly as drastic as we would have expected based on the sales volume chart. The hypothesis seems somewhat true but hard to prove that just from this visualization.

Location, Location, Location

The following visual is produced by aggregating sale price data based on neighborhood:



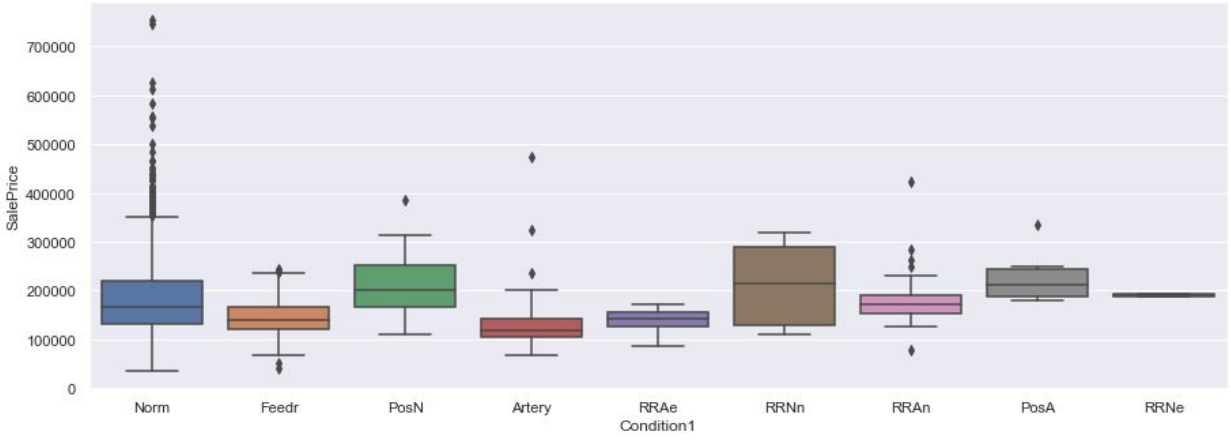
Seems that there is a lot of variance between the boxplots here which indicates that the neighborhood truly matters in determining sale price.

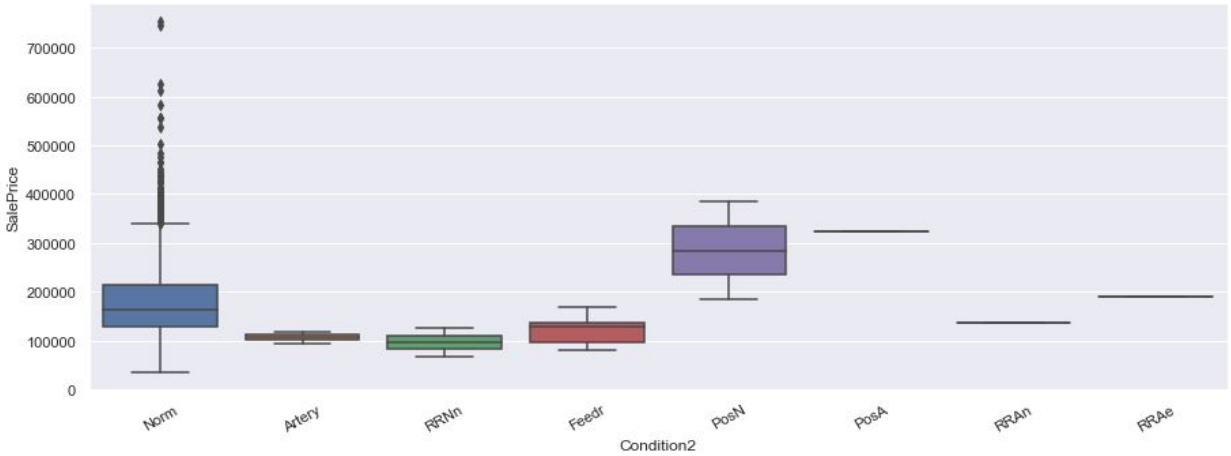
Other Features

Many other features could be affecting sale price. Let's not overlook those and share some additional visuals and observations.

Condition1, Condition2

We will take a look at Condition1 and Condition2. These are features which specify the proximity to certain city conditions (example PosN indicates a positive off-site feature including a park). My suspicion is that placement next to a positive feature such as a park will increase sales price of that home.

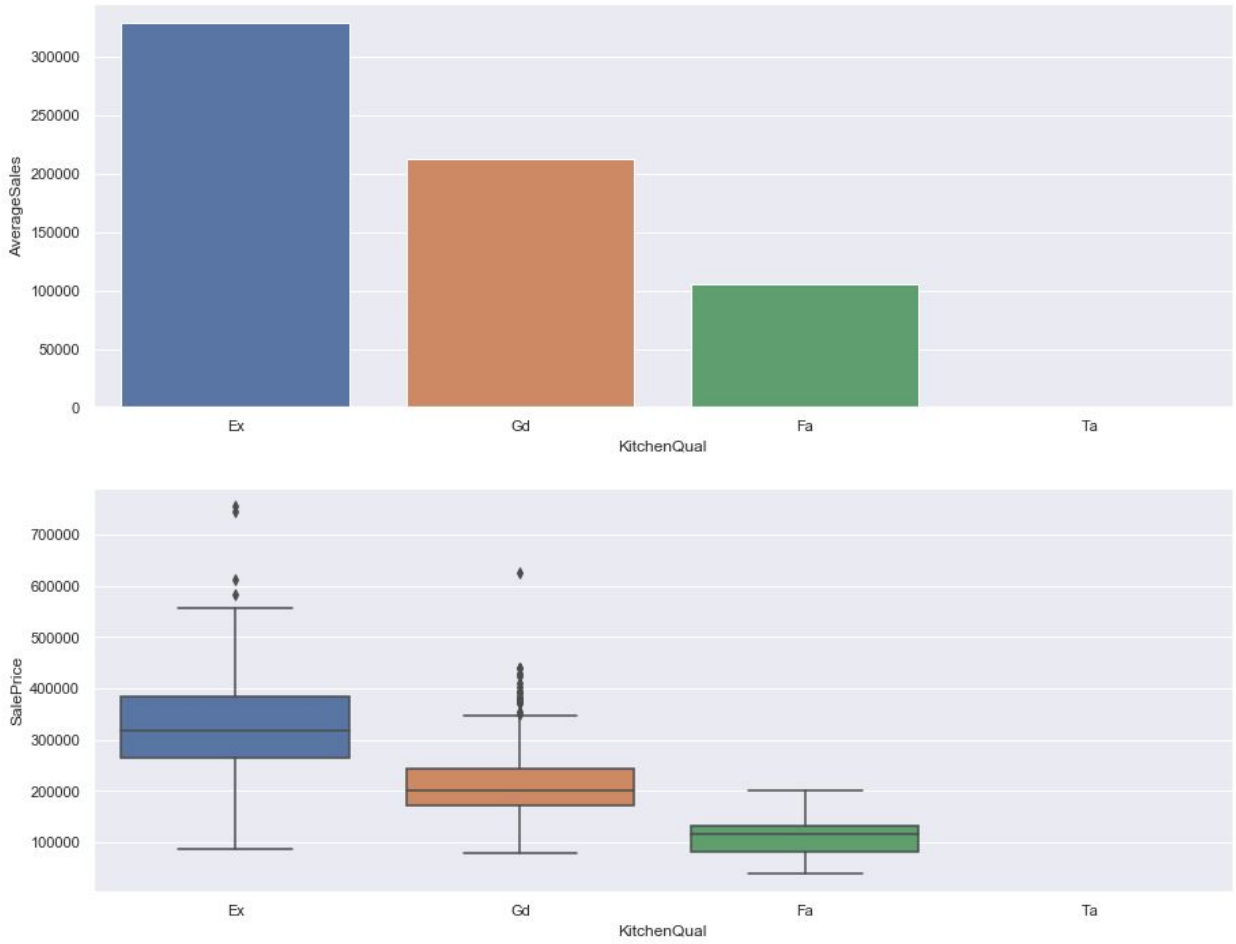




From the results this is clearly true! We can establish that having a positive off-site feature will directly influence sales price of that home.

Kitchen Quality

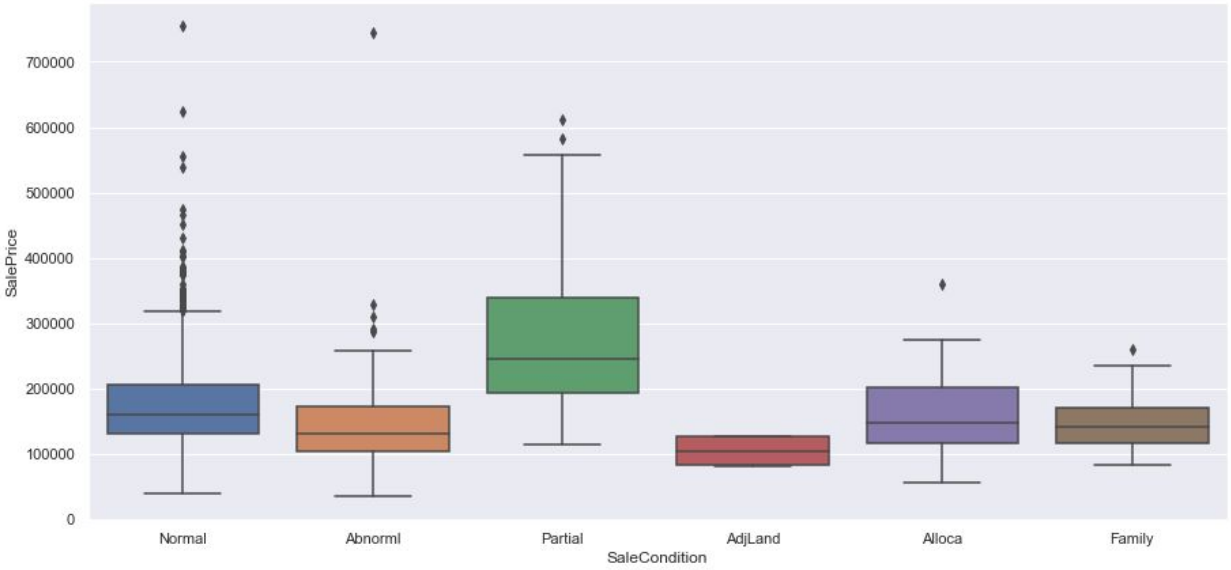
Let’s take a look at kitchen quality denoted by the 'KitchenQual' variable:



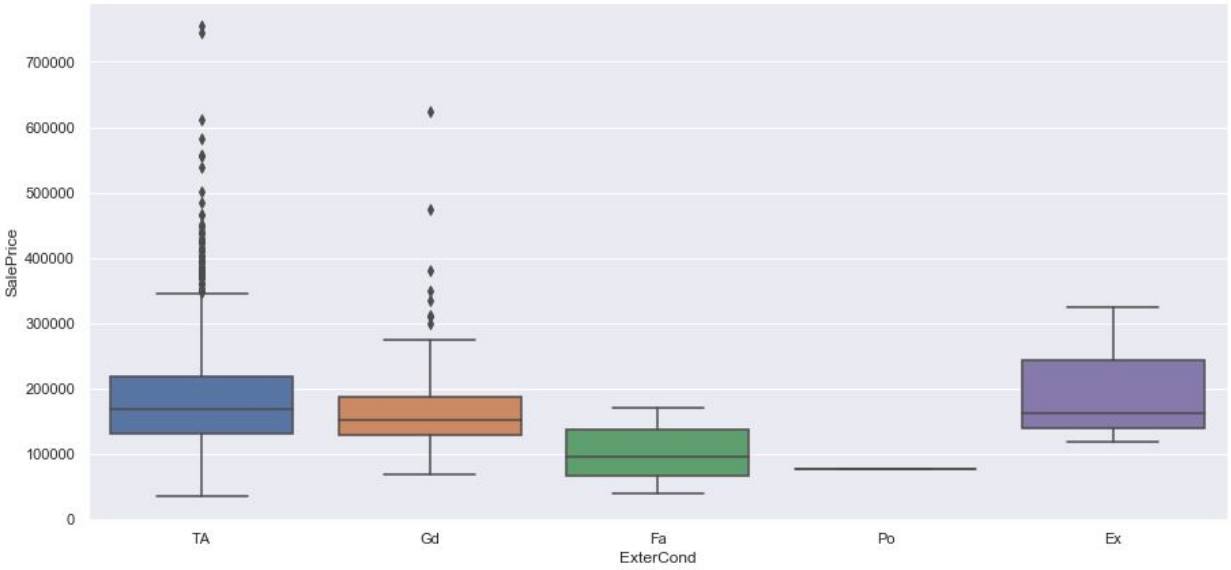
There is a clear relationship here between quality of a kitchen and the sales price. Its fairly obvious from these visuals that kitchen quality matters a lot when it comes to the final sale price of a home.

Sale Condition Feature

Let’s also examine the 'SaleCondition' feature. For this feature new homes are associated with the 'Partial' value. A hypothesis to test here would be that newer homes would fetch a higher sale price.

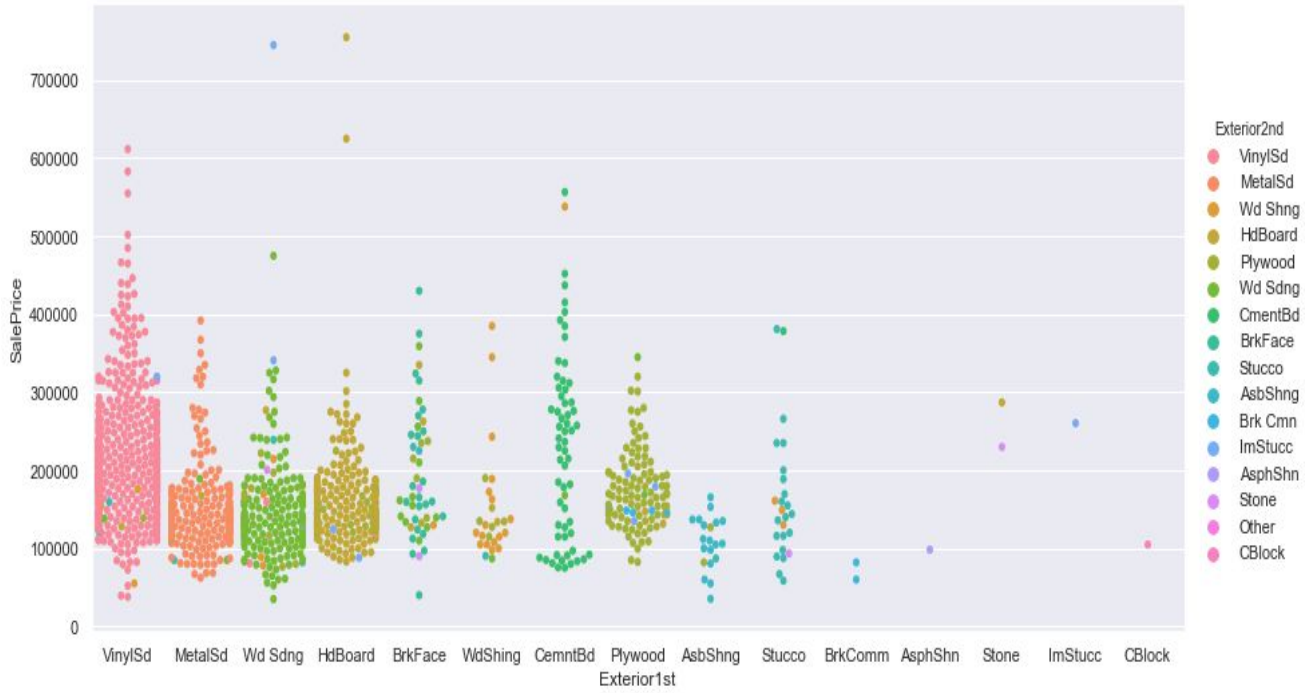


Exterior Condition



Exterior Material (1&2)

Since we are on the topic of exteriors, lets see if exterior material has anything to do with sale price. There are two variables which establish the exterior condition: 'Exterior1st' and 'Exterior2nd' (1st and 2nd exterior covering, respectively).

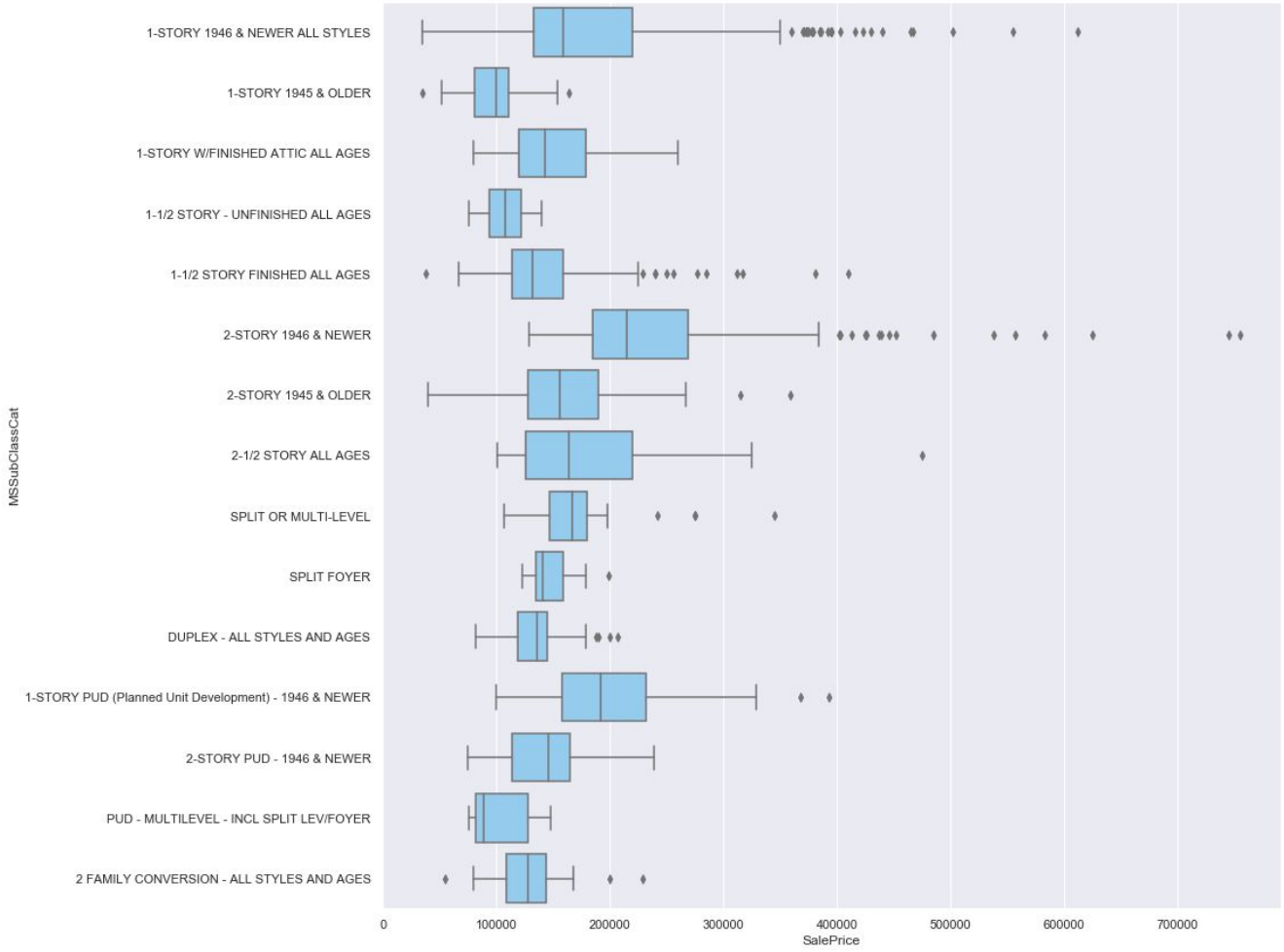


This is quite an interesting result. From the visual we can see that very rarely is there a secondary exterior covering of different type than the initial covering. That being said we can see a dominance of several exterior materials: Vinyl, Metal, Brick..etc.

From this visual we can also see that vinyl exteriors seem to dominate the range of sales prices greater than 200,000$.

Property Type: ‘MSSubClass’

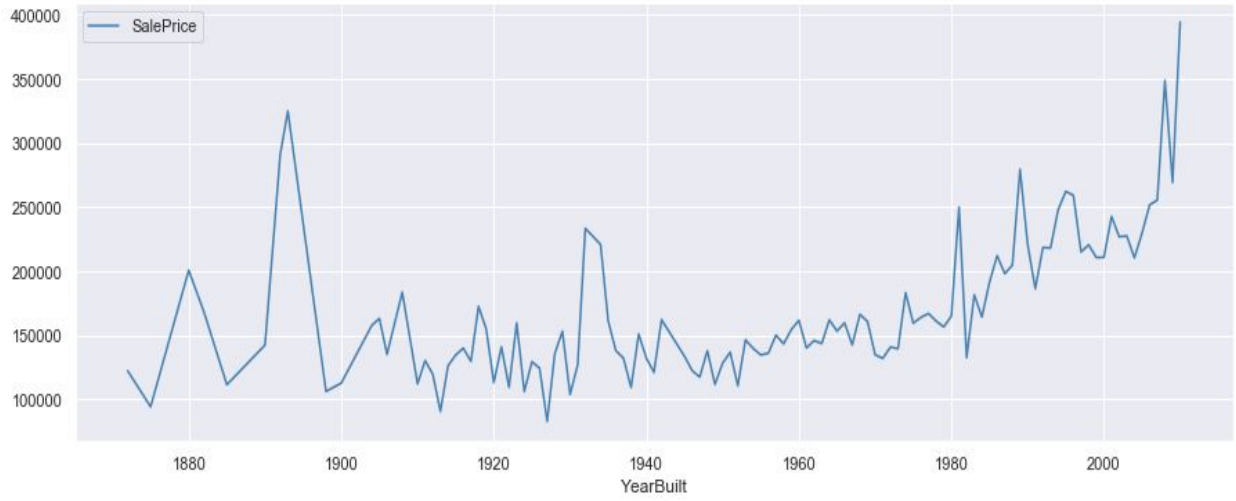
This property establishes the type of dwelling involved in the sale and has several classifications. Let’s take a look at these categories with respect to sale price:



These results are quite interesting since we see a highly varying distribution of Sale Price with respect to the MSSubClass field. This indicates a strong correlation between MSSubClass and sale price.

Year Built

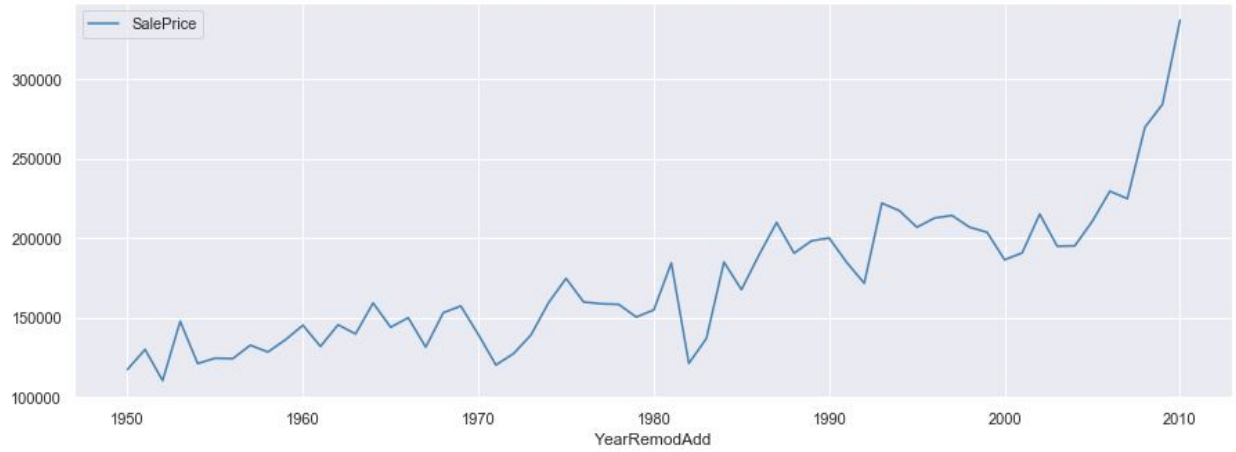
Let’s take a look at the relationship between year built ('YearBuilt') and sale price ('SalePrice'). The expectation is that the newer the home the more it will sell for. Let’s test this hypothesis:



The resultant graph shows that there is an observable trend. What’s interesting to note is that there is a large spike in value between 1890 to 1900. Perhaps this is due to some desirable architectural style which dominated that period. This may be worth looking into in the future. A bit of research on this time period may provide the explanation.

Year Remodeled

Let’s examine if a similar relationship can be observed between sale price and the year a remodel was added to the home (YearRemodAdd):



Conclusions

The results seem to indicate that square footage is the greatest influencer on home sale price. Locations seems to also have some influence on sale price yet its not as strongly correlated. For Ames, Iowa properties there is a very cyclical pattern to home sales with peaks in summer months. Attempts to correlate this to sale price were not successful (in spite of some pattern being observed, it seemed to be weak at most). The most revealing feature in this data seems to be TotalSF (total square footage of the home). This parameter was created by combining livable square footage, basement square footage and garage area. When examining its correlation to sale price, we observe it is the highest of all features.

It’s very clear from this data that success in predicting home price comes from looking at many of these features. There seem to be correlations everywhere and with over 79 variables, it’s difficult to say that one is more valuable than the other. I believe the solution to making an accurate prediction will be to consider all of these variables in order to establish some type of statistical model for home price..