Introduction

The Dataset

Pricing a home can be very difficult and currently real estate agencies rely on manually comparing a home to several other comparable homes (also known as ‘comps’) sold recently in the same vicinity. This process yields less than ideal results since the agent will likely only choose a few homes and a few features causing the process to be prone to error. Therefore a more data-driven approach may be the solution to understanding this process.

The dataset being used is from a Kaggle competition and contains 79 variables detailing almost every feature of residential homes in Ames, Iowa. It can be found below:

Home Sale Prices - Ames, Iowa

The dataset contains about 1459 rows of data including the following features: Sales price, Building Class, Zoning, Lot Frontage, Lot Area, Street, Alley, LotShape, LandContour, Utilities, LotConfig, LandSlope, Neighborhood, Year Built.. Etc.

Initial Assumptions about the data

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.

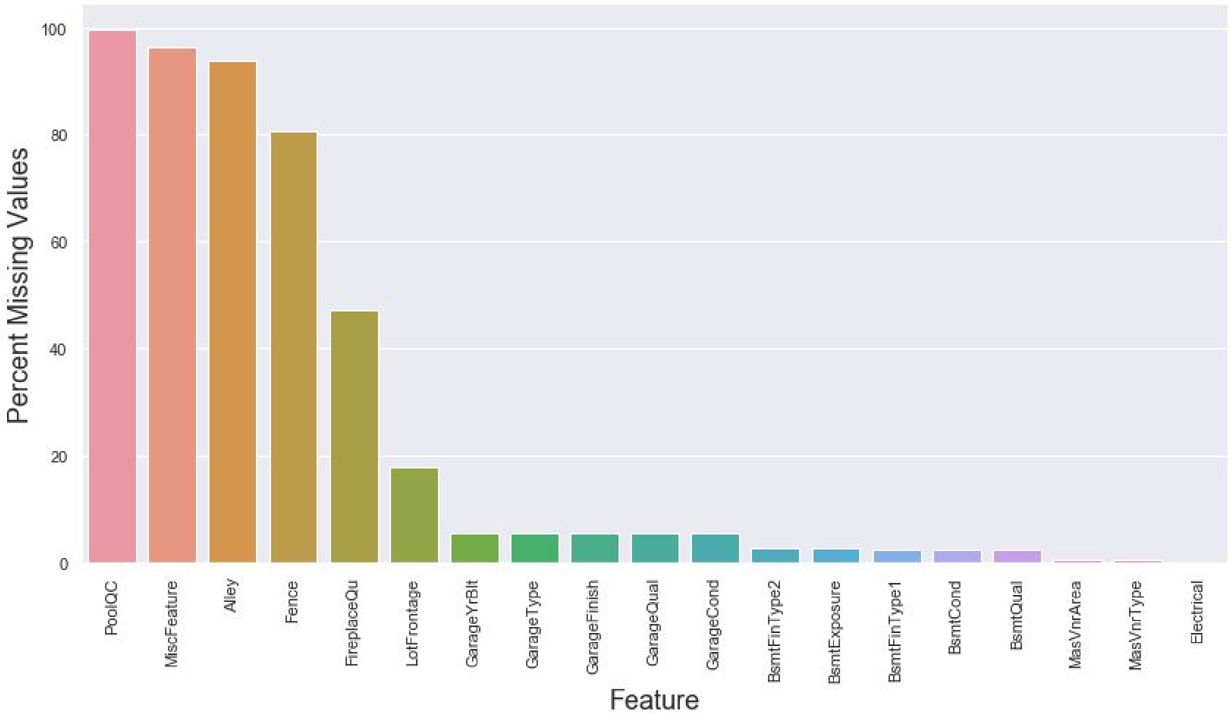
The other variable of interest here will be the square footage or overall size of the home. In this dataset there are many variables which describe the home size. These include: Lot area, Livable square footage, Basement square footage, Garage area in square feet and several others.

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.

Data Wrangling Steps

Cleaning the Data

First off we will perform some visualizations to determine how much data is missing from this dataset. We obtain the following result showing percentage of missing data by feature:



The fields **PoolQC**, **MiscFeature**, **Alley**, **Fence** and **FireplaceQu** all have Na values which correspond to absence of that feature. For example ‘Na’ for **PoolQC** means the property has no pool. To simplify the data processing here we would like to convert these values to a representative string, in this case we will use "None".

For square footage values such as **LotFrontage** (square footage on the street connected to the house), its highly unlikely to get a value of 0. To treat these missing values we will replace them with the median **LotFrontage** value.

For **GarageType**, **GarageQual** and **GarageCond** we will replace missing data with 'No Garage' as indicated by the data dictionary.

For **BsmtQual**, **BsmtCond**, **BsmtExposure**, **BsmtFinType1** and **BsmtFinType2** we can replace the null values with the string "No Basement".

For **MasVnrArea** and **MasVnrType** here null values indicate no veneer applied to the home. So let’s fill in 0/None for those respectively:

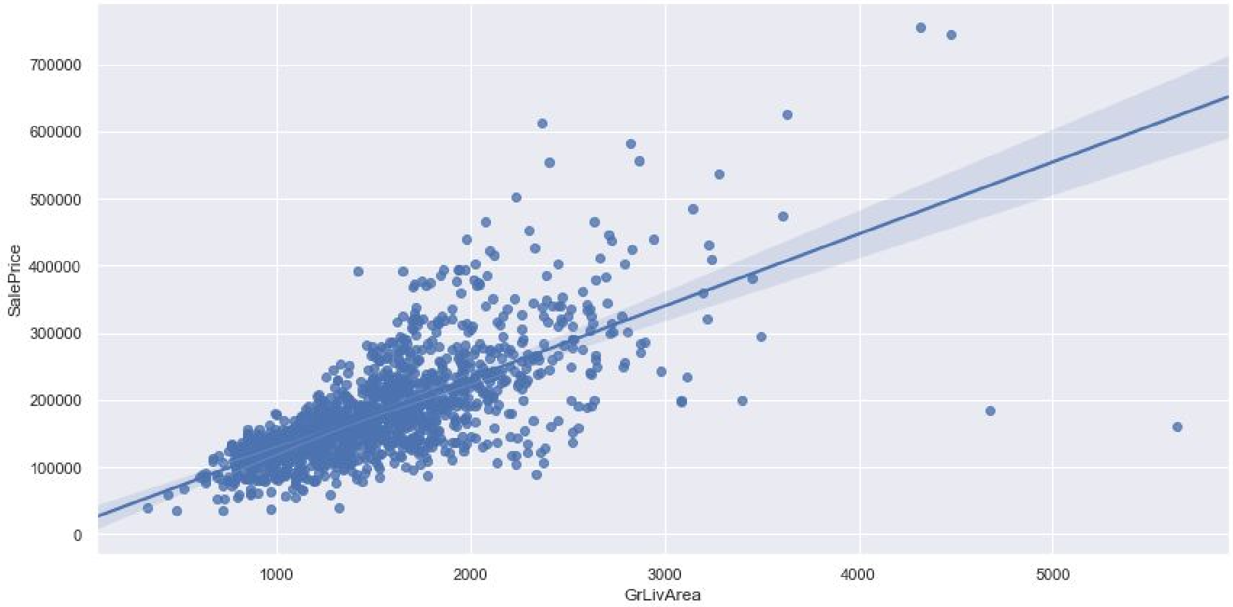
For **Electrical** we can replace with mode since it is categorical data (and only 1 missing value is present).

Finally for **MSSubClass** we can replace null values with "No Building Class" as suggested by the data dictionary.

After performing these substitutions we now have a complete dataset with all missing parameters replaced with more meaningful values.

Dealing with Outliers

The dataset’s author mentions that there are 4 outliers in the dataset. Specifically, anything greater than 4000 square feet living area should be removed from the dataset. Before eliminating these, let’s take a look at these data points in a visualization showing **GrLivArea** (the total living area of the house) versus the sales price of that home (**SalePrice**):



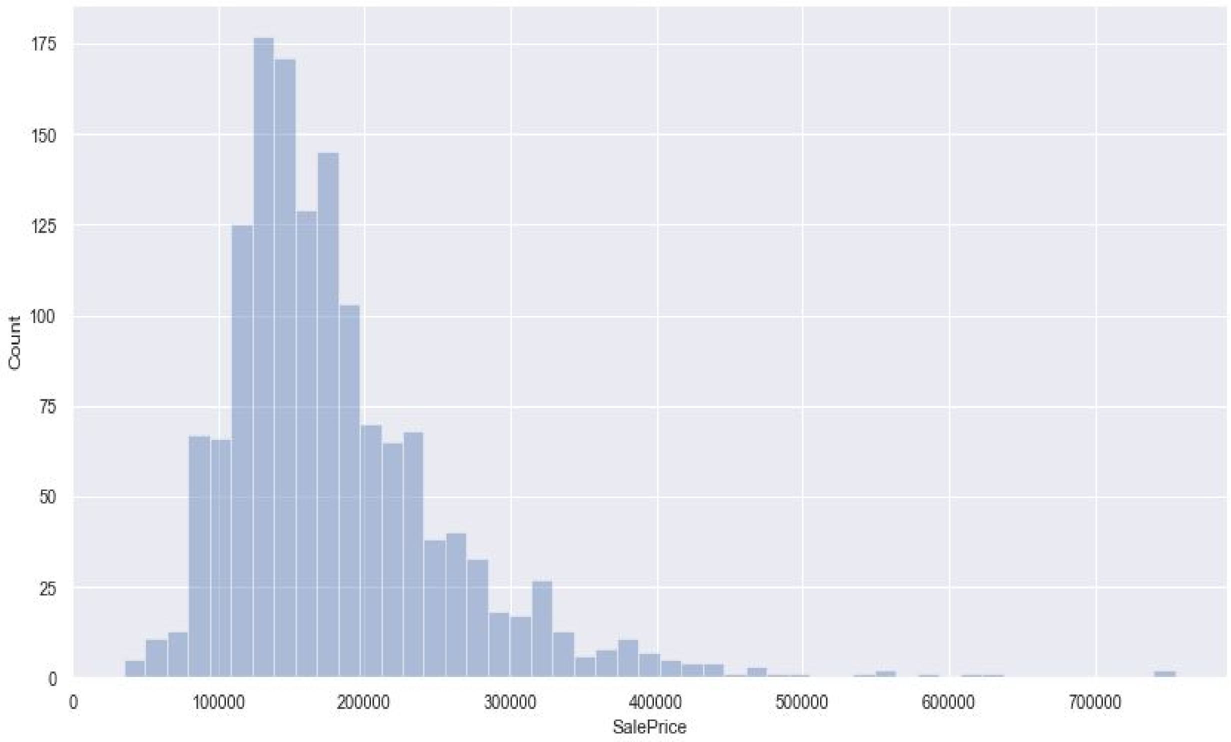
The four data points in question can be identified fairly clearly in this visualization. Lets isolate these data points:

idx	SalePrice	GrLivArea
523	184750	4676
691	755000	4316
1182	745000	4476
1298	160000	5642

Index 691 and 1182 might actually be realistic data points since they were sold at a very high price point. For now we will leave these outliers in the dataset. For modeling purposes we do not want to assume a perfectly linear relationship so we will re evaluate whether or not to keep these in our dataset based on modeling results. Simply based on this graph it appears the two data points below 200,000 sale price could be true outliers.

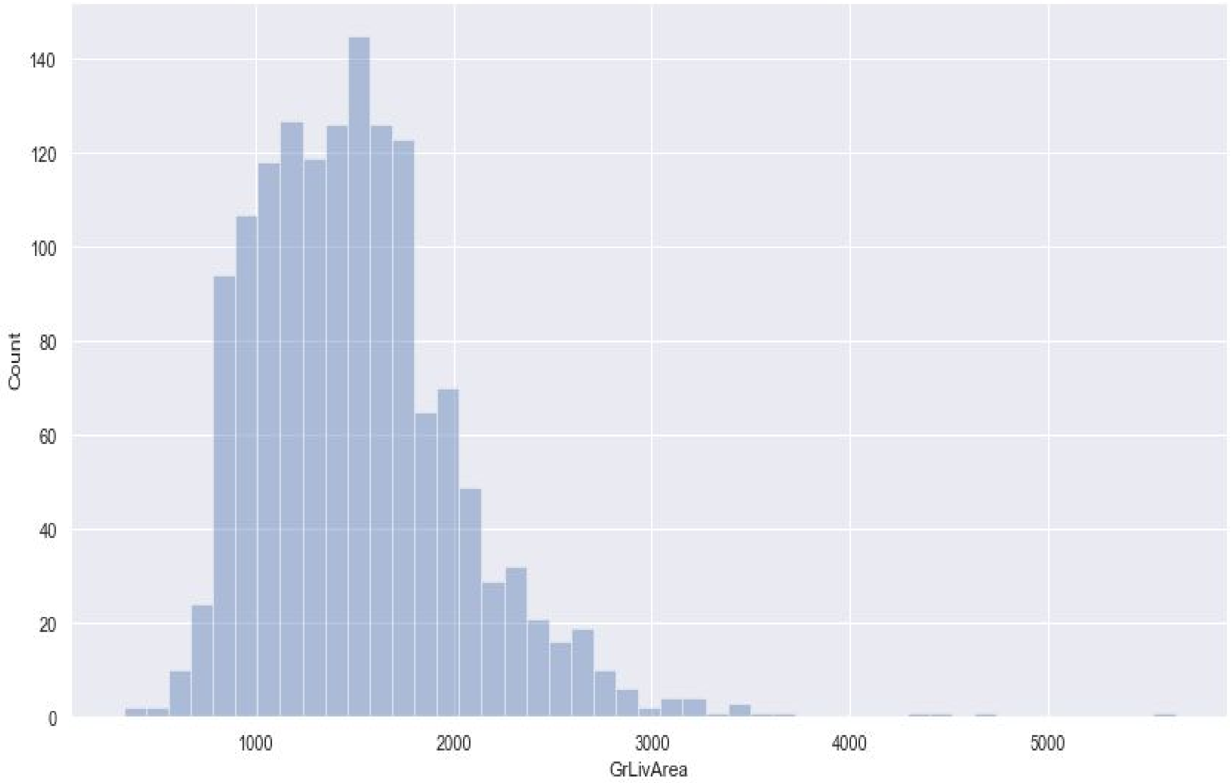
Data Exploration

Now that we have cleaned the data, lets begin exploring it and seeing if we can confirm some of the initial assumptions made in the introduction. First we will plot the distribution of the **SalePrice** variable:



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?

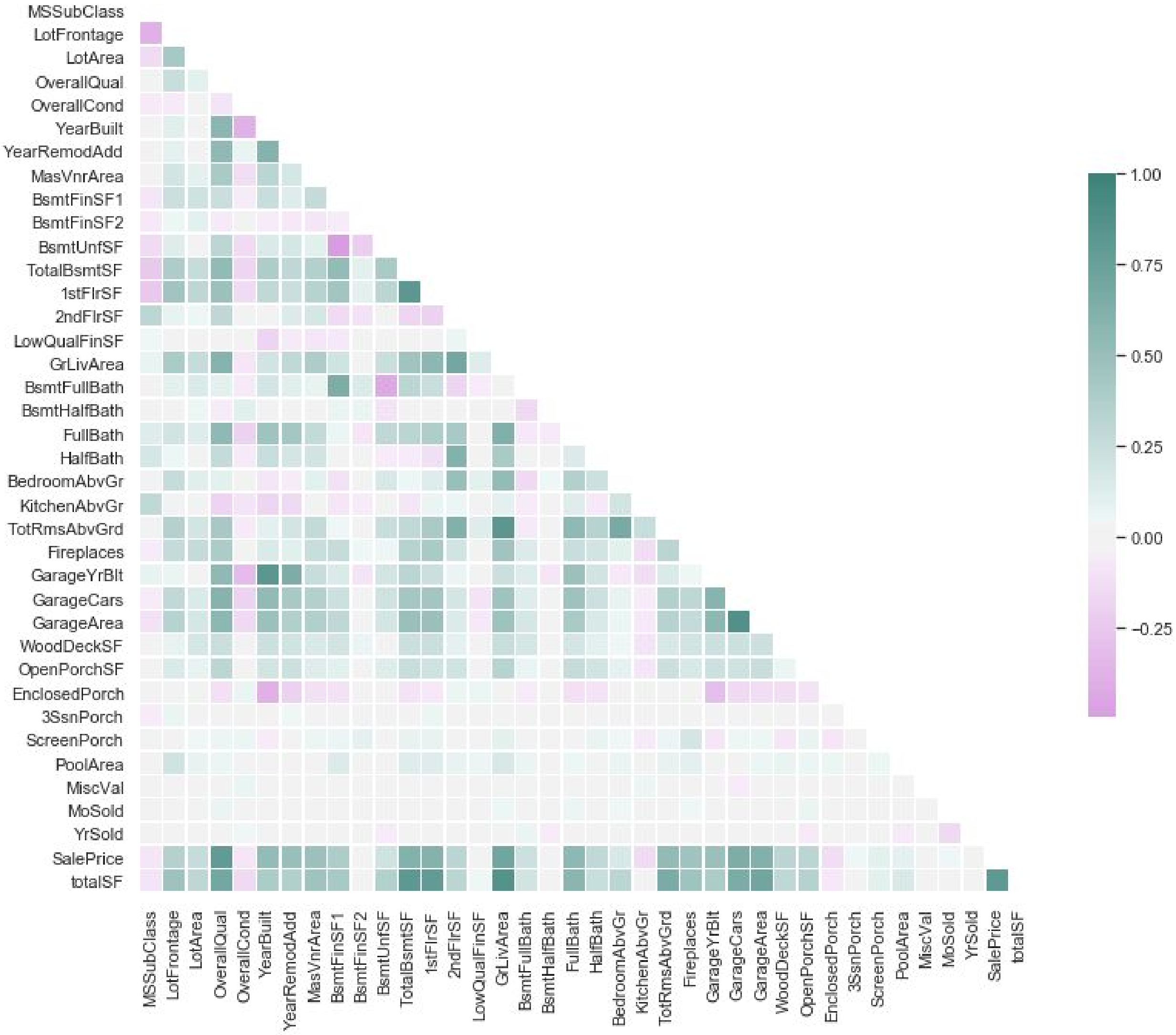
Let's take a look at what seems to be the most obvious influencers on sales price, livable square footage (the variable **GrLivArea**). The goal here is to see if we see a similar distribution to that of sales price. If so, this would indicate perhaps that this skewed sales price distribution might in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



This is interesting, it looks like the living area (**GrLivArea**) 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:

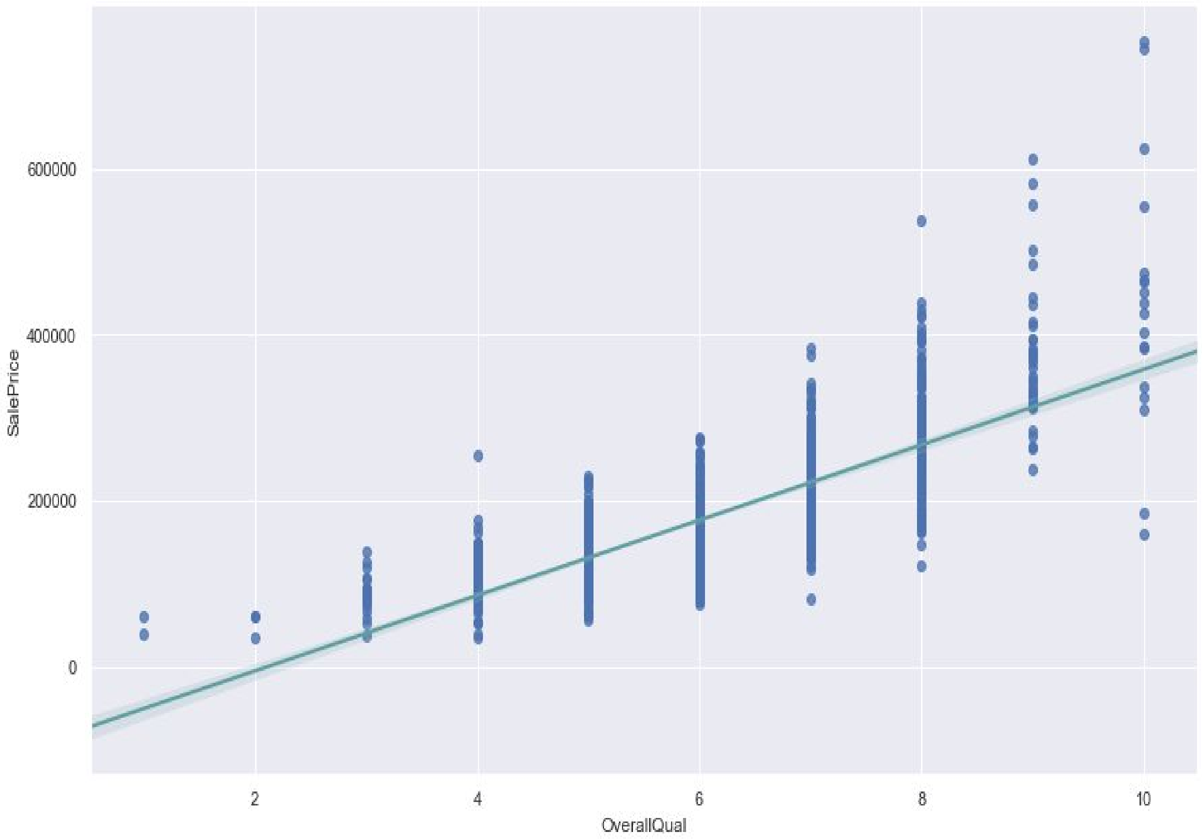


The parameter with highest correlation to sale price is in fact **OverallQual**. However we do see that right behind this feature are **GrLivArea** (livable space in square feet), **TotalBsmtSF** (square footage of the basement) and **GarageArea**. So this seems to support the initial hypothesis that square footage has the largest impact on sale price.

Examining Numerical Features with High Correlation to Sale Price

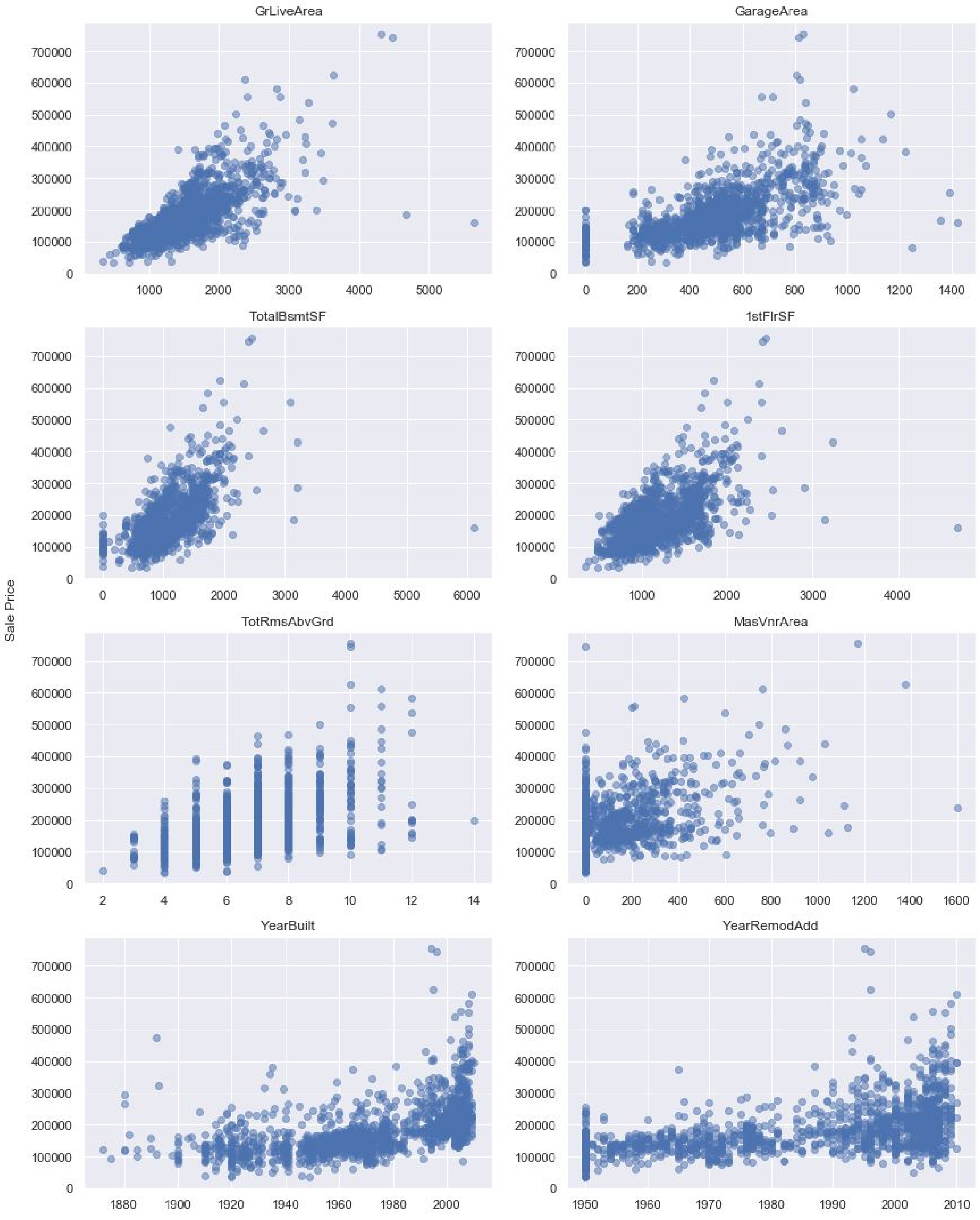
Overall Quality of Homes

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 or **OverallQual** (ranging from 1-10) we see a fairly wide range of sale prices. However there is a clear trend here. Via the 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:

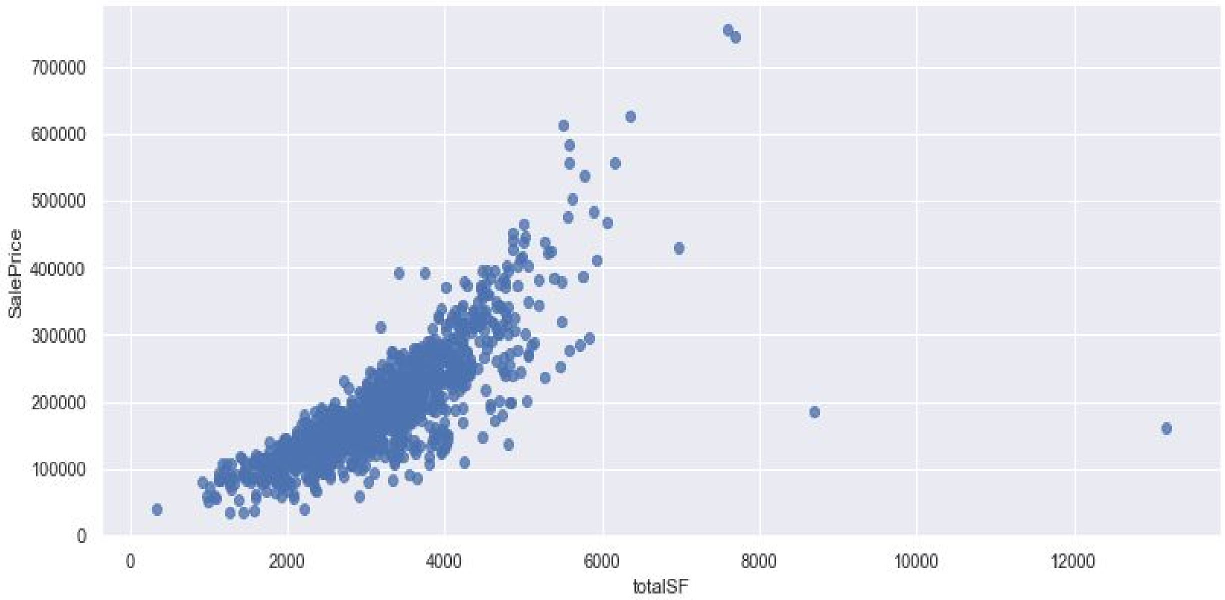

Sale Price
GrLiveArea
0
100000
200000
300000
400000
500000
600000
700000
1000
2000
3000
4000
5000
GarageArea
0
100000
200000
300000
400000
500000
600000
700000
0
200
400
600
800
1000
1200
1400
TotalBsmtSF
0
100000
200000
300000
400000
500000
600000
700000
0
1000
2000
3000
4000
5000
6000
1stFlrSF
0
100000
200000
300000
400000
500000
600000
700000
1000
2000
3000
4000
TotRmsAbvGrd
0
100000
200000
300000
400000
500000
600000
700000
2
4
6
8
10
12
14
MasVnrArea
0
100000
200000
300000
400000
500000
600000
700000
0
200
400
600
800
1000
1200
1400
1600
YearBuilt
0
100000
200000
300000
400000
500000
600000
700000
1880
1900
1920
1940
1960
1980
2000
YearRemodAdd
0
100000
200000
300000
400000
500000
600000
700000
1950
1960
1970
1980
1990
2000
2010

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.

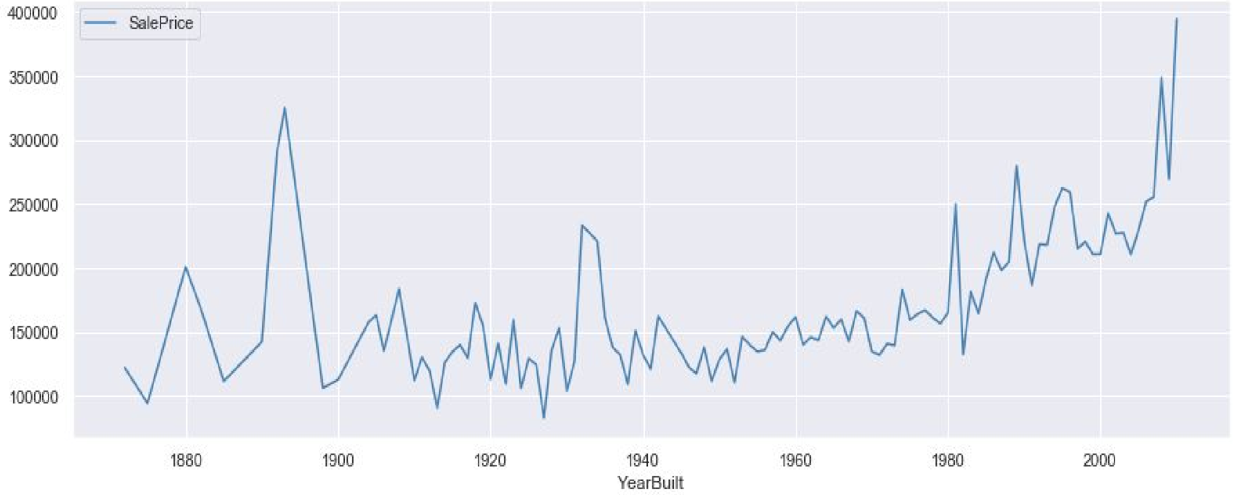
What’s even more interesting here is that if we re evaluate the **Pearson correlation coefficients** with the new **totalSF** parameter in the datasets, we obtain the following result:

$$\boldsymbol{P_{totalSF}\,(.807) > P_{OverallQual}\,(.791)}$$

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

Year Built

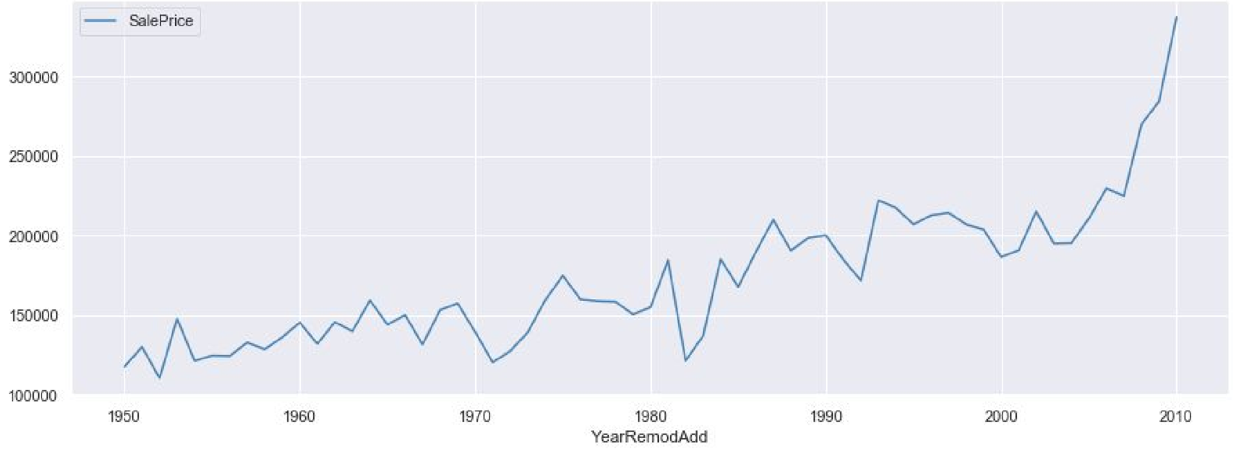
Let's take a look at the relationship between year built (**YearBuilt**) and sale price (**SalePrice**). The expectation is that the newer the home the more it will sell for. Let's test this hypothesis:



The resultant graph shows that there is an observable trend. What's interesting to note is that there is a **large spike in sale price 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.

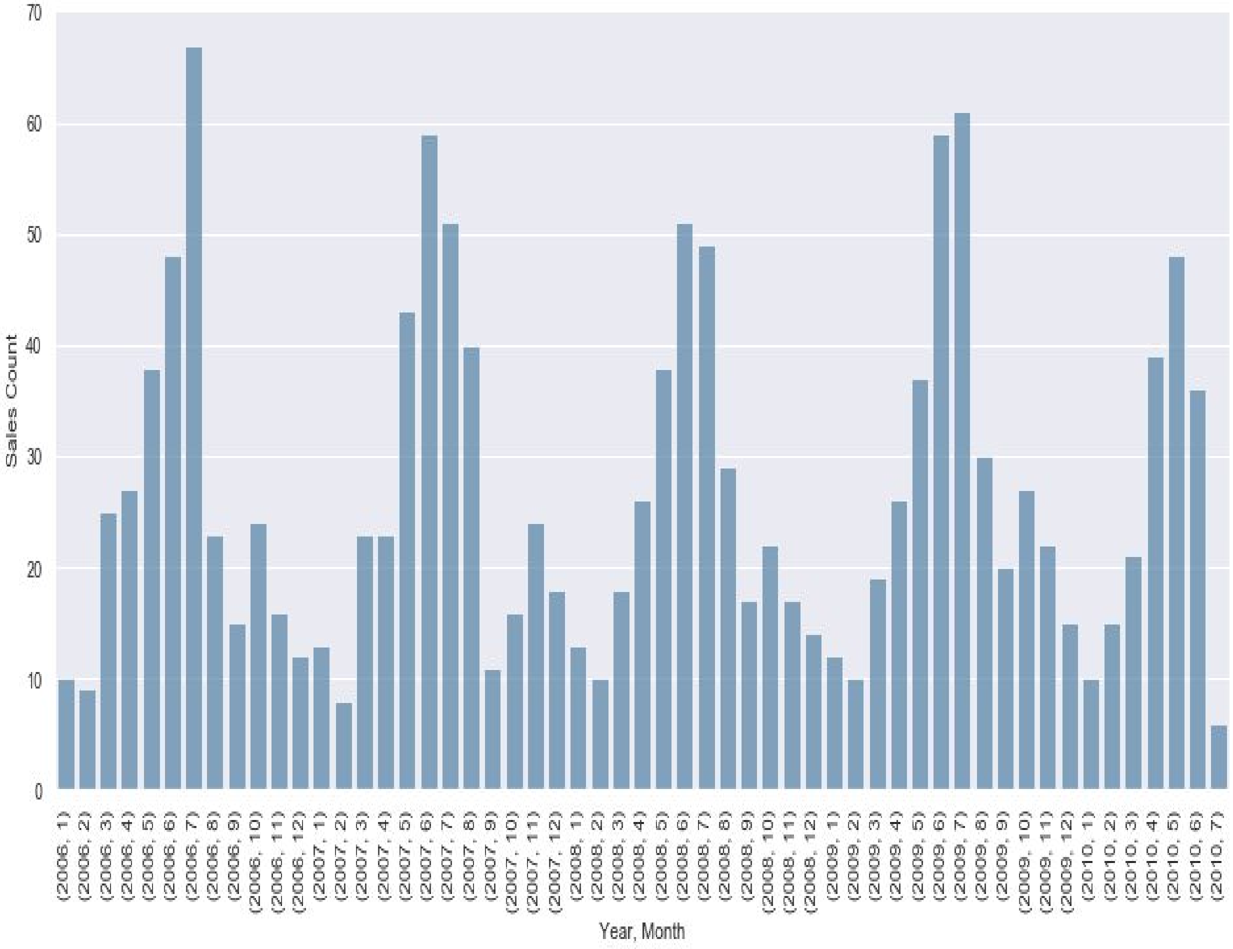
Year Remodeled

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

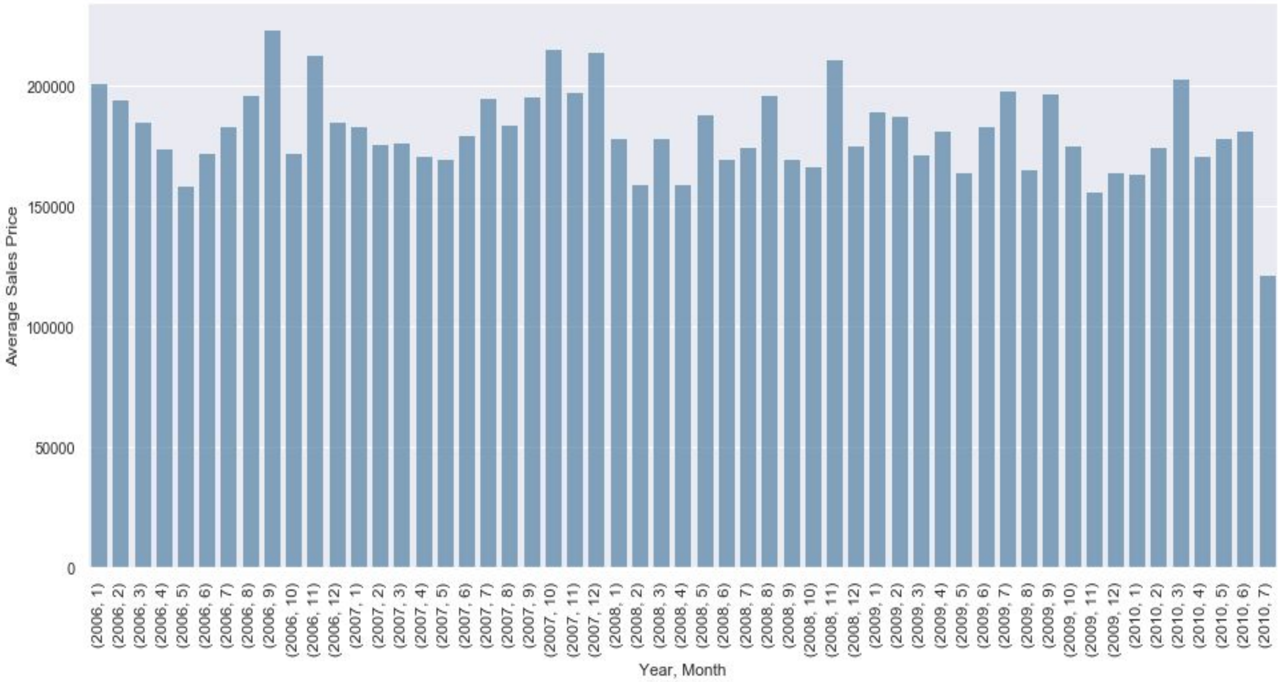


Evaluating Sales Volume

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:

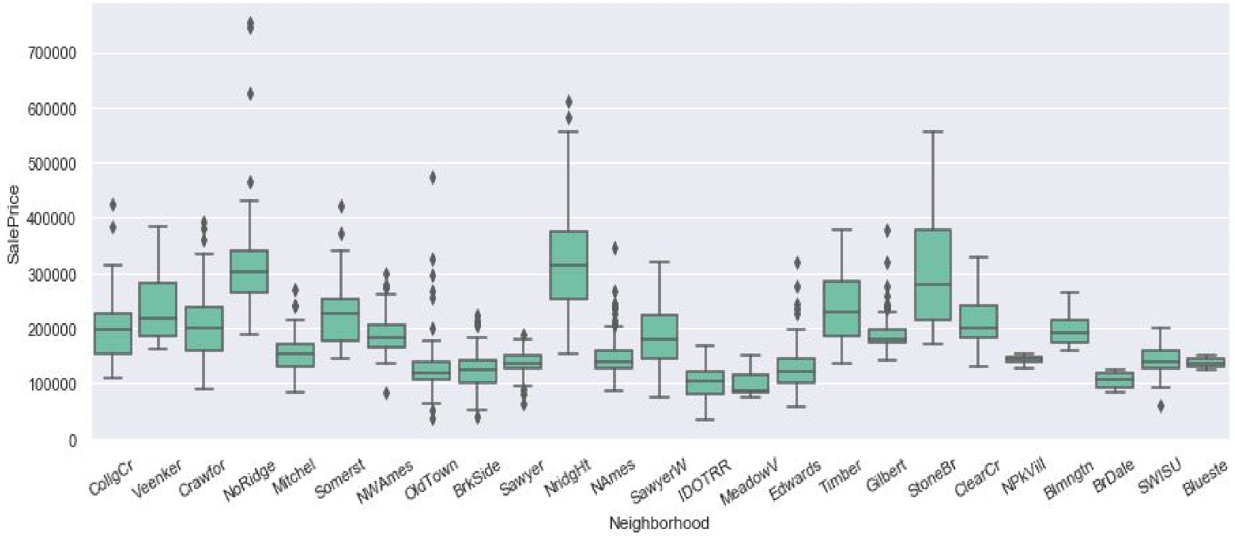


From this result there seems to be a trend, but not nearly as drastic as we would have expected based on the sales volume chart. The hypothesis seems somewhat true but hard to prove that just from this visualization.

Categorical Features

Location, Location, Location

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

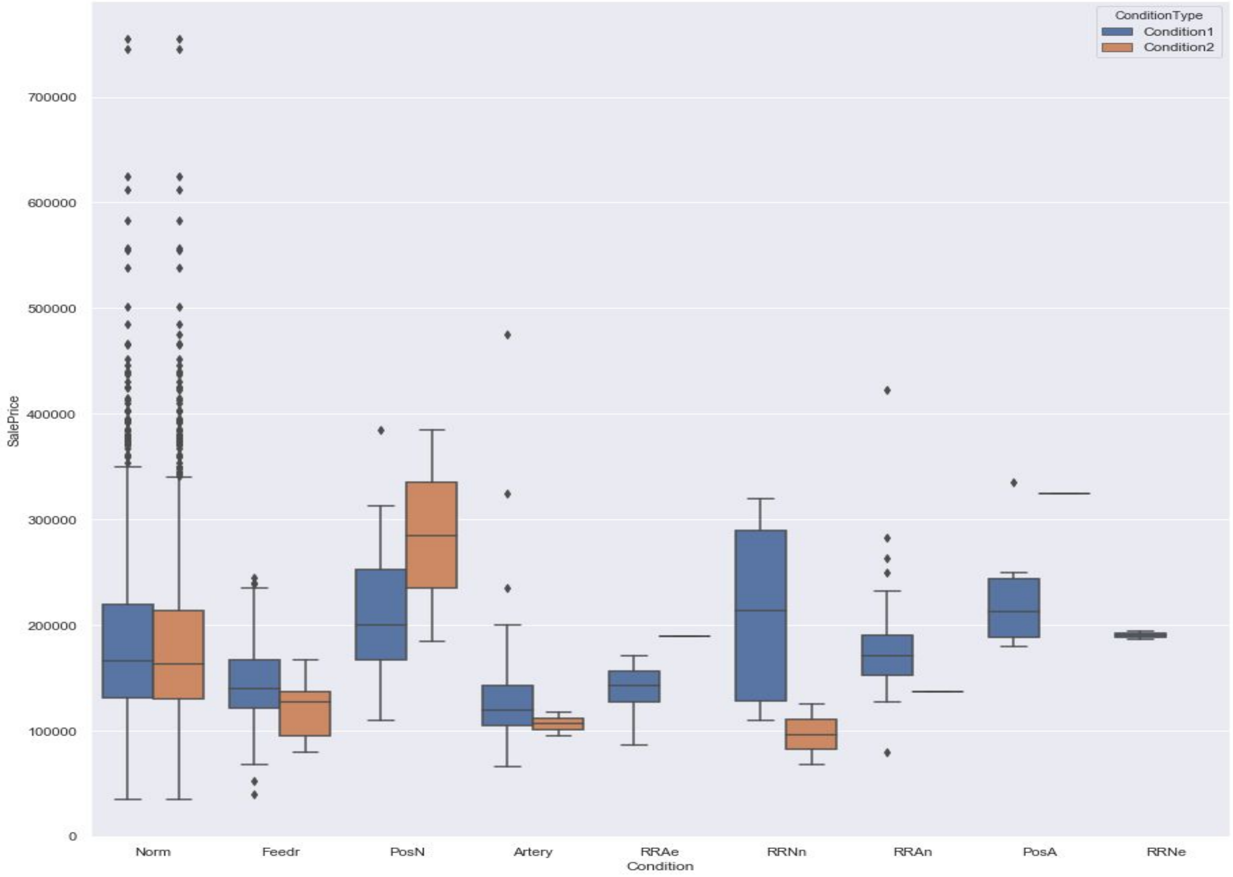


Seems that there is a lot of variance between the boxplots here which indicates that the neighborhood truly matters in determining sale price. This is confirmation of the initial assumption that neighborhood location within Ames, Iowa will be relevant.

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

Condition1, Condition2

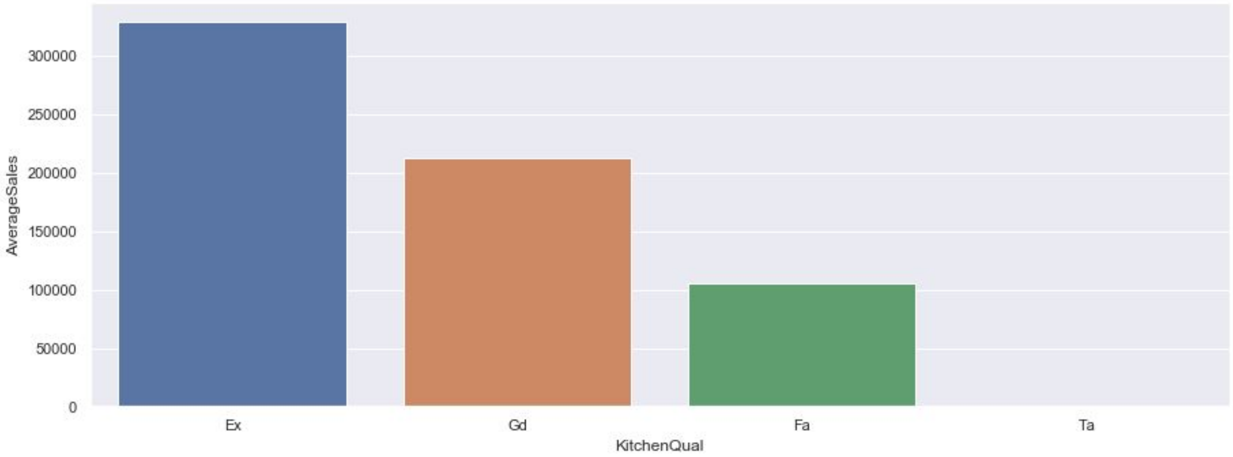
We will take a look at **Condition1** and **Condition2**. These are features which specify the proximity to certain city conditions (example PosN indicates a positive off-site feature including a park). My suspicion is that placement next to a positive feature such as a park will increase sales price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

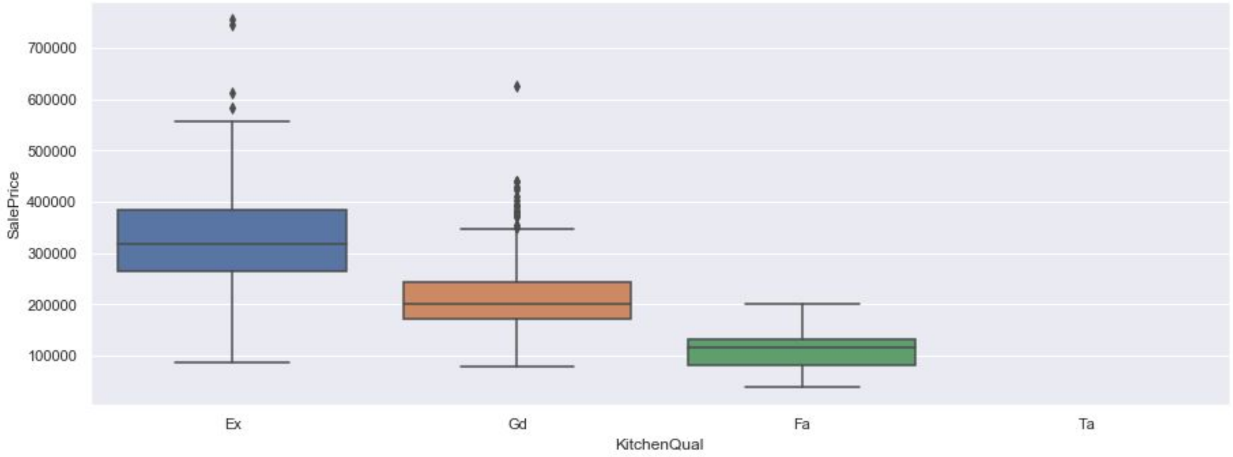


From the results this is clearly true! We can establish that having a positive off-site feature will directly influence sales price of that home.

Kitchen Quality

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

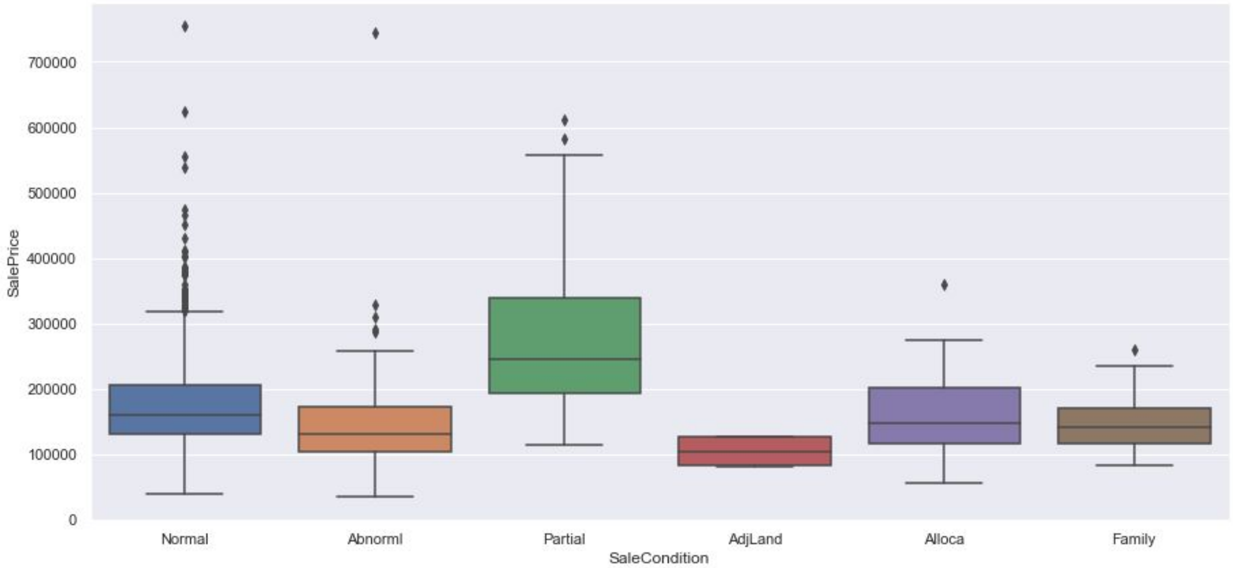




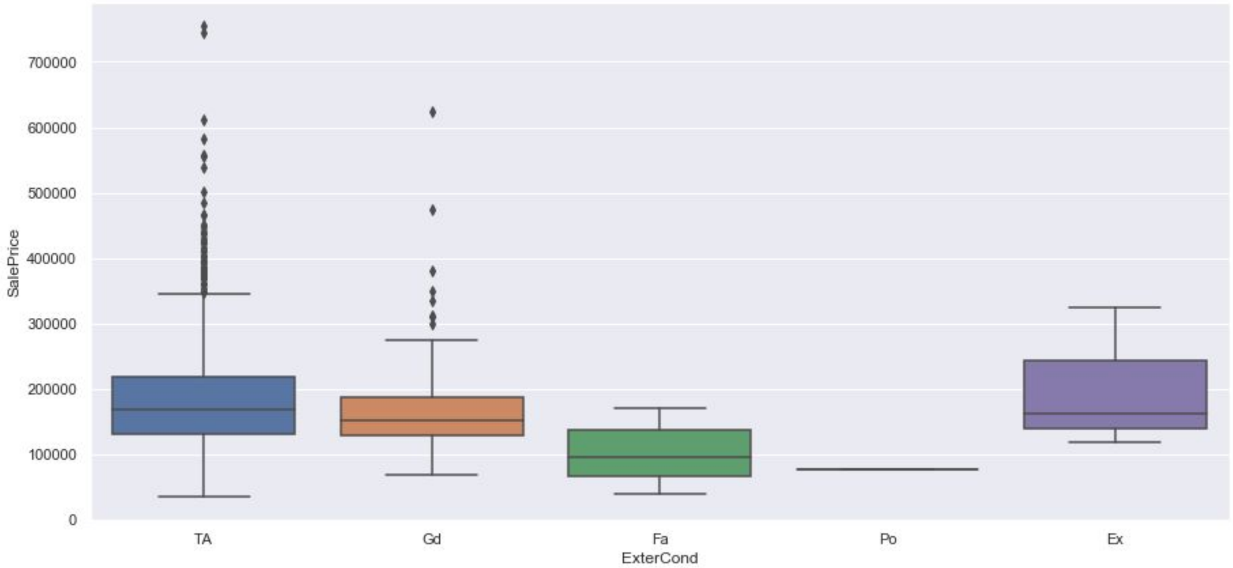
There is a clear relationship here between quality of a kitchen and the sales price. Its fairly obvious from these visuals that kitchen quality matters a lot when it comes to the final sale price of a home.

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.

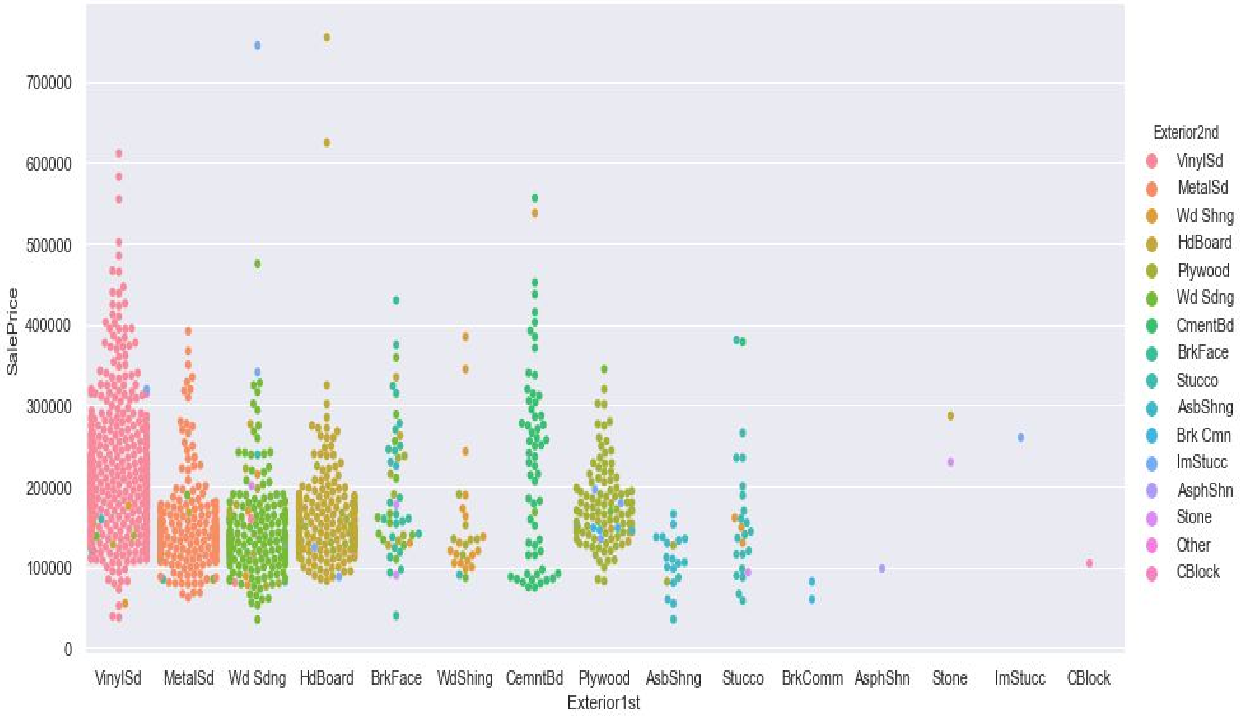


Exterior Condition



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).

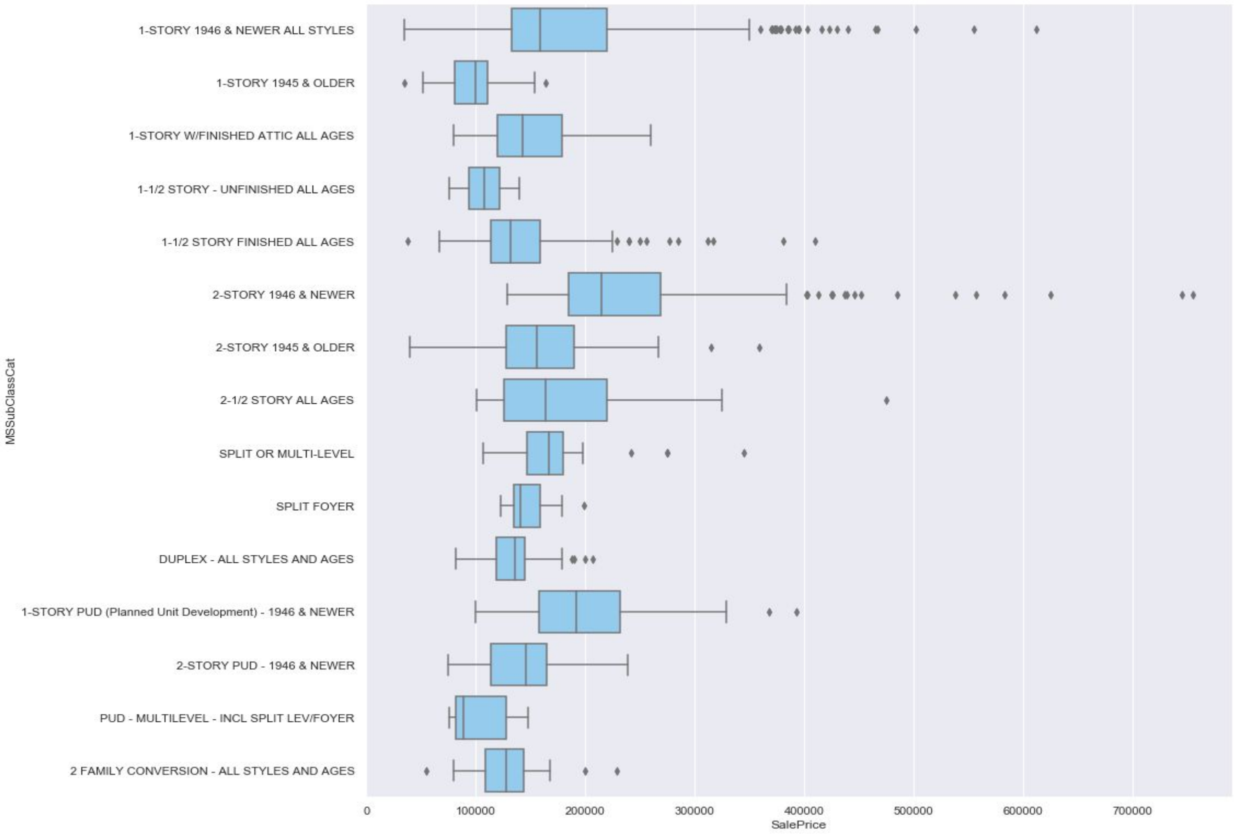


This is quite an interesting result. From the visual we can see that very rarely is there a secondary exterior covering of different type than the initial covering. That being said we can see a dominance of several exterior materials: Vinyl, Metal, Brick..etc.

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 feature 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 **SalePrice** with respect to the **MSSubClass** field.

EDA Summary of Results

The results seem to indicate that square footage is the greatest influencer on home sale price. Location or **Neighborhood** seems to also have some influence on sale price yet it’s 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.

The significance of this correlation is however questionable and should be analyzed further using some inferential statistics techniques. Additionally some inferential statistics will be applied to determine how significantly prices differ between each neighborhood.

Inferential Statistics

While exploring the Ames, Iowa dataset (see previous section) several interesting discoveries were made. Some exploratory data analysis revealed strong correlations between the target variable **SalePrice** and some of the dataset's features. More specifically this correlation with **SalePrice** was observed with the Overall Quality score and several square footage related parameters (**GrLivArea**, **TotalBsmtSF** and **GarageArea**), with **OverallQual** being the highest. Later in the analysis the square footage parameters were combined to create a new feature **totalSF**. When combined this feature had a Pearson coefficient greater than **OverallQual** (with respect to SalePrice). This difference however was minimal at roughly .016 so this is worthy of some statistical exploration.

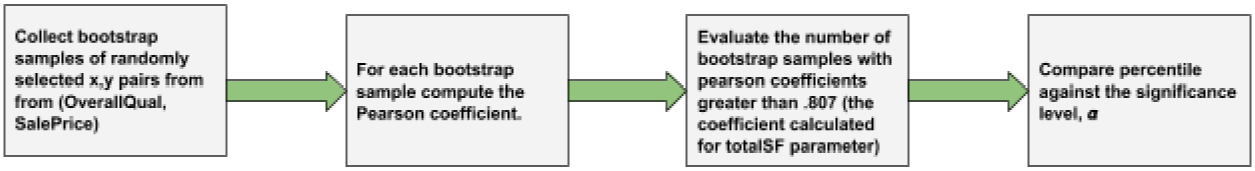
While exploring the categorical data, a categorical box plot showed that **Neighborhood** seemed to have quite a large impact on **SalePrice**. This is worth exploring, in particular determining which neighborhoods have statistically significant differences in mean sale prices. To do this we will perform some hypothesis tests comparing the mean sale's prices between neighborhoods.

Exploring totalSF and OverallQual Features

The hypothesis to explore here is whether the difference in the Pearson coefficients of the **totalSF** and **OverallQual** features is statistically significant. Can we say that **totalSF** has a stronger correlation to **SalePrice**? Or is this difference negligible? To test this difference we will perform some bootstrap hypothesis testing. For this test, the following null and alternate hypotheses will be used:

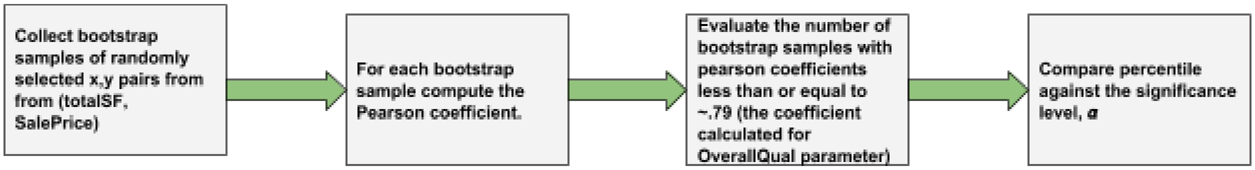
> **Ho:** The difference in correlation coefficients between **totalSF** and **OverallQual** is negligible.
>
> **Ha:** The **totalSF** parameter has a greater correlation to **SalePrice**
>
> ***α*****:** .05

The process for this test will be as follows:


Collect bootstrap samples of randomly selected x,y pairs from from (OverallQual, SalePrice)
For each bootstrap sample compute the Pearson coefficient.
Evaluate the number of bootstrap samples with pearson coefficients greater than .807 (the coefficient calculated for totalSF parameter)
Compare percentile against the significance level, α

Results: The results show that **5%** of replicates generated had pearson coefficients greater than or equal to .8075185 (the coefficient for **totalSF**).

With an alpha of .05 we would normally accept our null hypothesis since **p>=α,** however in this case we are exactly on the threshold so this decision becomes quite difficult. What happens when we look at this the other way around?


Collect bootstrap samples of randomly selected x,y pairs from from (totalSF, SalePrice)
For each bootstrap sample compute the Pearson coefficient.
Evaluate the number of bootstrap samples with pearson coefficients less than or equal to ~.79 (the coefficient calculated for OverallQual parameter)
Compare percentile against the significance level, α

In other words let's run the same bootstrap test taking replicates instead from the **totalSF** data and compare these to the pearson coefficient calculated for **OverallQual**.

Results: The results for this test differ significantly from the first showing that **28.86%** of the replicates have coefficients calculated to be .79098 or lower. Given this result we are led to accept the null hypothesis that the difference between correlations is negligible.

Finally using scipy.stats in python we compute the coefficients and their relative p-values with the following result:e

totalSF pearson:

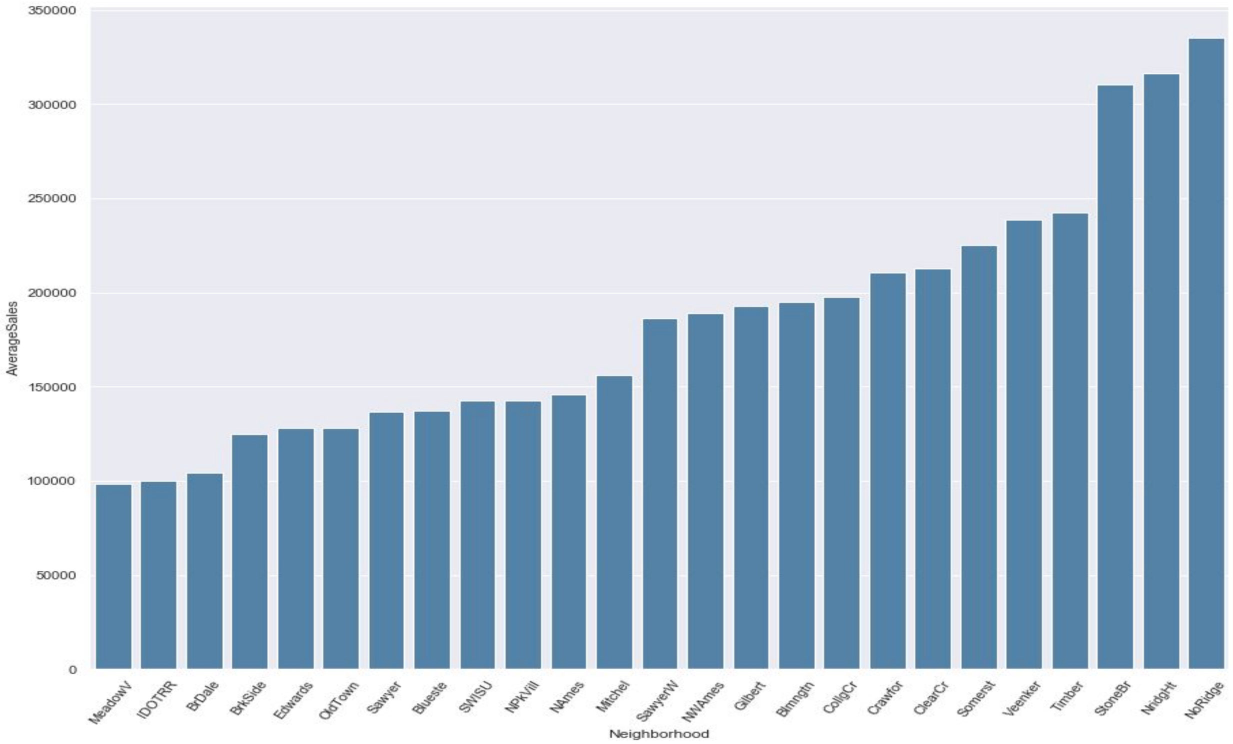
(***P***= 0.8075184760515013, **p-value** = 0.0)

OverallQual pearson:

(***P***=0.7909816005838051, **p-value** = 2.185675268e-313)

Examining Differences Between Mean Sale Price of Neighborhoods

To begin this exploration let's first visualize the difference between mean sale price of neighborhoods:



Seems like there are three neighborhoods with average sales significantly higher than the rest.

Let's test statistical significance between neighborhoods. For this test we will look at the lowest average sales **MeadowV** versus the neighborhood with highest average sales **Timber** (excluding the 3 potential outliers StoneBr, NridgeHt, NoRidge).

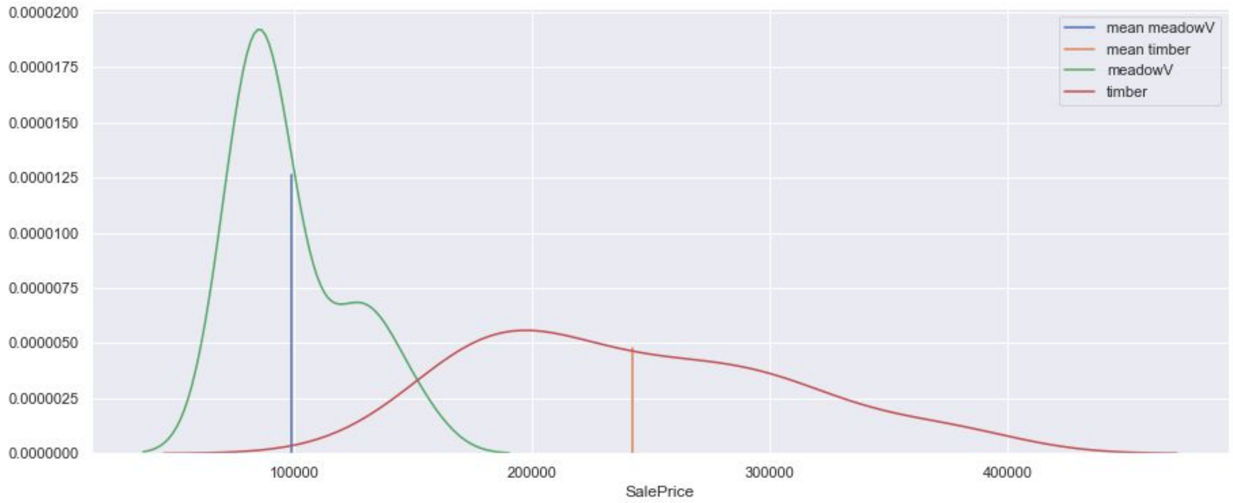
We will set up this test as follows:

Ho = There is no difference between mean of MeadowV and Timber home sale prices.

Ha = The means are not equal (two tailed test)

Alpha = Lets use an alpha value of .05 or 5%

First lets plot the individual distributions to see how they differ visually:

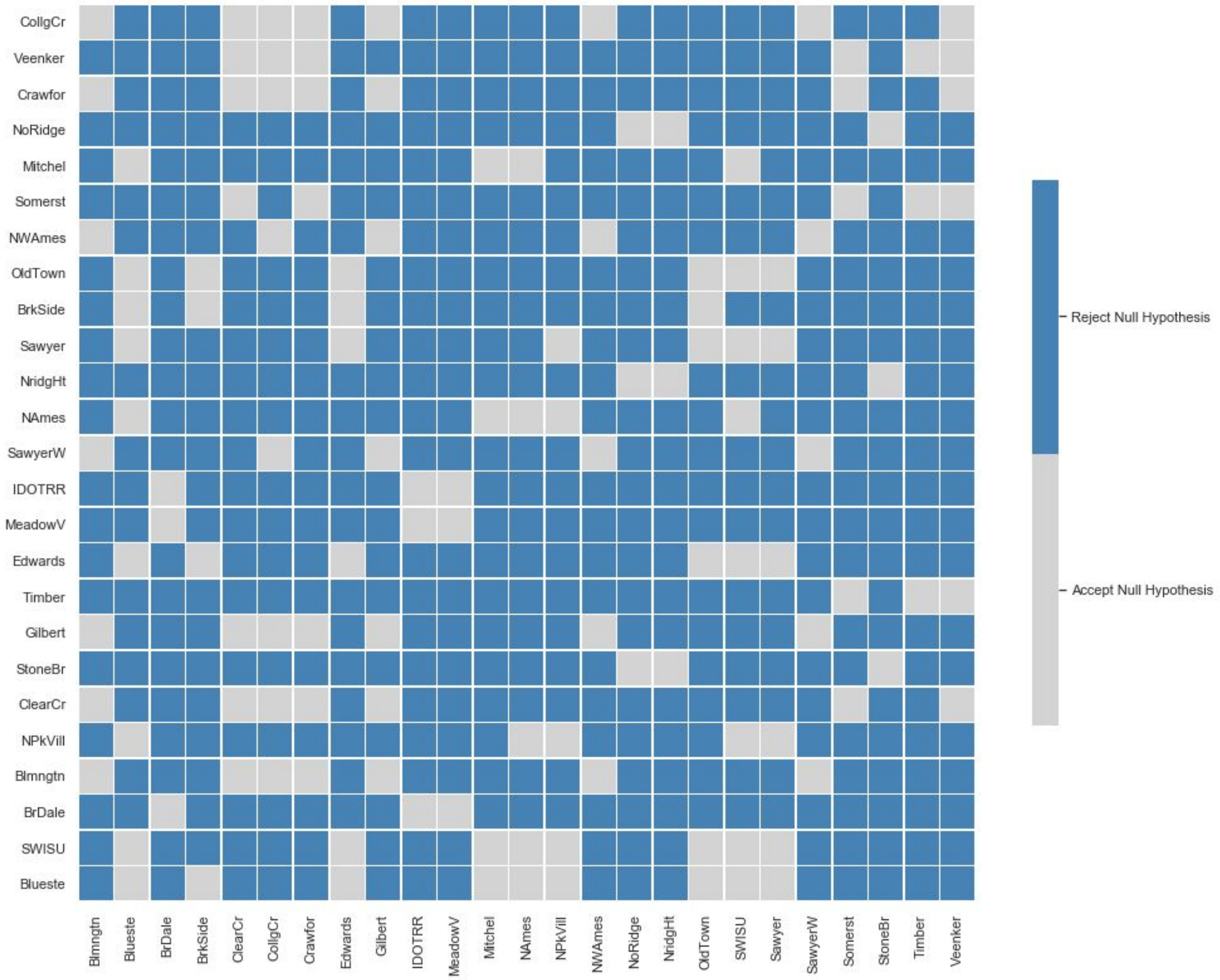


Quite a large difference in distributions. Next we perform some calculations to confirm what we are observing visually:

Standard Error calculated = 11759.927
Z-value calculated = -12.217 (standard deviations from the mean)
P-value obtained = 2.522 e-34

So from the result it's fairly clear that these means differ significantly. This leads us to reject the null hypothesis that the means are equal in favor of the alternate. What about other neighborhoods? Let's consider all neighborhoods and evaluate statistical significance for each against the others.

To do this we construct an array of all possible comparisons. Essentially this is an array of neighborhood pairs to be used for comparison. We then perform the same analysis shown above for each pair of neighborhoods. For these comparisons we are using the same alpha value, null and alternate hypotheses. Rather than listing each case, a visual has been prepared (using seaborn heatmap function) to summarize the results of these comparisons:

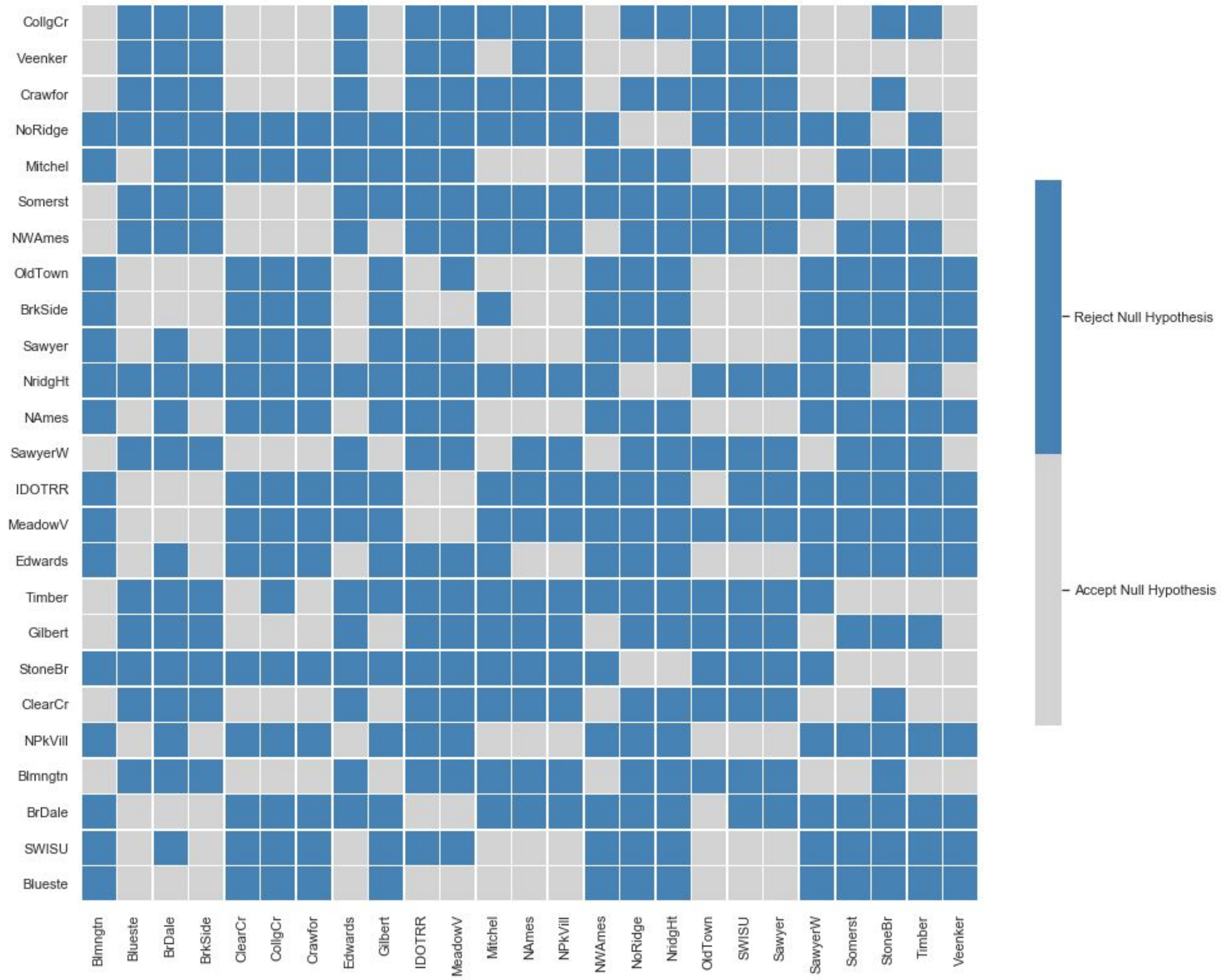


It’s clear from this visual that many test cases are being performed here. As a result of this with an alpha of .05 we have a 5% chance of incurring a Type I error (where we reject a true null hypothesis). This means that out of 625 comparisons about 31-32 results will be subject to Type I errors.

To correct for this we need to apply a more conservative approach, in this case we will use the Bonferroni method. This method states that the alpha value used should be divided by the number of comparisons being made. So in this case:

$$\boldsymbol{\alpha} = .05/625 = .00008 \text{ or } .008\ \%$$

So rerunning the comparisons with this new and much more conservative alpha value, we obtain the following results:



In comparison, **494** statistically significant mean differences were found before the Bonferroni method was applied, whereas **414** were observed after applying this method.

This seems to show that the mean differences in many cases are quite drastic since even with such a small alpha value of .00008 a majority of test cases still yielded the same result.

Conclusions

It's very clear from this data that success in predicting home price comes from looking at many of these features. From the data exploration and statistical analysis it seems fairly apparent that neighborhood, square footage and overall quality of homes are all significant contributors to a home's sale price. I believe the solution to making an accurate prediction will be to consider all of these variables in order to establish some type of machine learning model for home price.