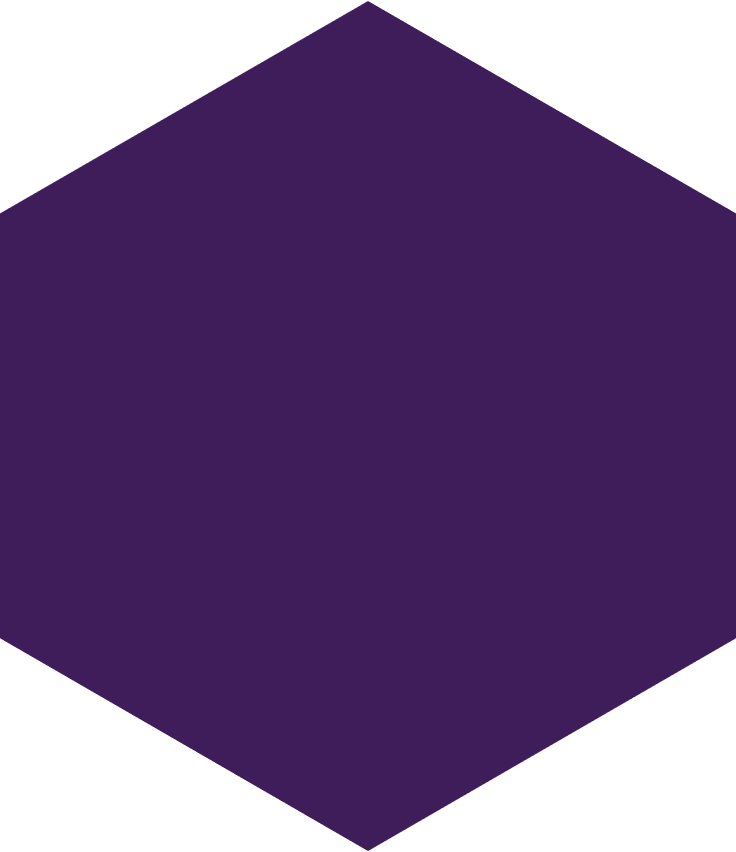


|  |
| --- |
| **Prediction of Home Sale Price** |
| By: Leoman Momoh |
| A detailed report on construction of a Home Sale Price prediction model using python code, machine learning algorithms, and Statistical analysis. |
|  |



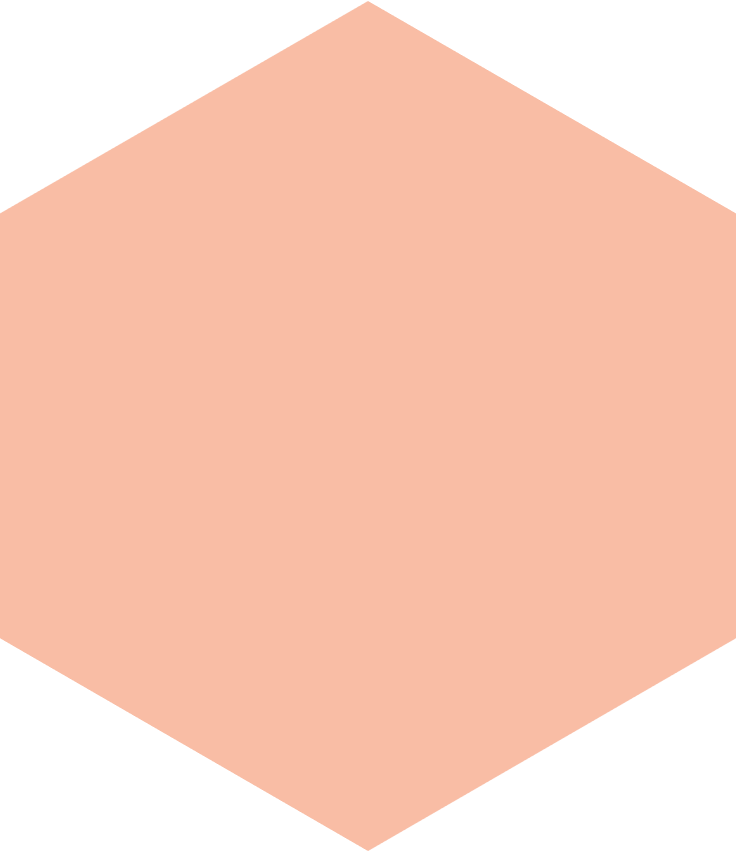


Table of Contents

[The Hard Truth 2](file:///C:\Users\eshai\SB%20Projects\Cap%202\Deliverable%20Report.docx#_Toc508333610)

[Introduction: 2](file:///C:\Users\eshai\SB%20Projects\Cap%202\Deliverable%20Report.docx#_Toc508333611)

[The problem and data available: 2](file:///C:\Users\eshai\SB%20Projects\Cap%202\Deliverable%20Report.docx#_Toc508333612)

[Deliverables: 2](file:///C:\Users\eshai\SB%20Projects\Cap%202\Deliverable%20Report.docx#_Toc508333613)

[Exploratory Data Analysis 3](#_Toc508333614)

[Statistical Inference 6](#_Toc508333615)

[Preprocessing and Model Construction 9](#_Toc508333616)

[1st Model 9](#_Toc508333617)

[2nd Model 10](#_Toc508333618)

[3rd Model and Boosting 11](#_Toc508333619)

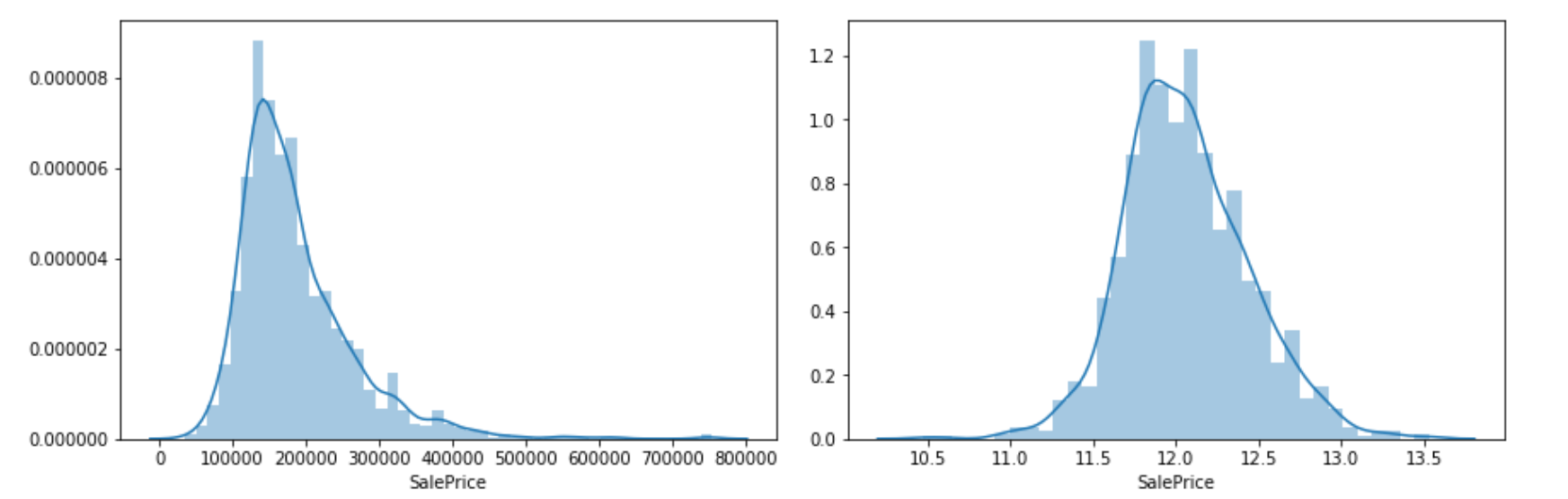
[Key Insights 12](#_Toc508333620)

|  |  |  |
| --- | --- | --- |
| A peak at the problem  What’s the Right Price? |  | **The Hard Truth** • • • |
| **?**  “Any buyer can make a mistake, and if they don't keep a level head, that mistake can end up costing them a lot of money,"  -Herman Chan  (San Francisco Real Estate Broker) |
| Introduction: According to Zillow (a leading real estate company), median home prices are projected to rise in the U.S. market. Since home prices are predicted to rise, no buyer wants to pay for a home that is worth less than the listed price. Many buyers rely on realtors to find the home with the best value. Unfortunately, sometimes buyers can be left with an overvalued home once the realtor has claimed his or her commission on the sale. The problem and data available: Buyers need a tool to estimate what the actual sales price of a home should be. With an estimation tool, buyers won’t have to rely solely on the word of a realtor or a seller that a home is appropriately priced. Using a data set available from Kaggle, I aim to give buyers a resource to determine what the actual price of a home should be based on its features.  The data set I plan to use for my analysis contains data on houses in Ames, Iowa. The data set contains 79 distinct feature variables and missing values that will have to be either omitted or filled. I plan to use statistical analysis, python programming, and machine learning algorithms to identify key features that best explain the variation between sale prices. Deliverables: I plan on delivering a housing price prediction model that will be beneficial to buyers, sellers, and investors. In addition to the model, I plan to provide a report that provides details on my process as well as the features that have the greatest influence on price. Buyers, sellers, and real estate investors will find this model useful when determining what the price of a home should be. Buyers and sellers could use this model as a guide to haggle on the price of a home leading to savings or profit on the home depending on the party. Real estate investors could use the insight from the report to focus on investing in homes with valuable features. In conclusion, my aim for this project is to assist all parties by providing a mathematical formula to determine what the starting price for a home should be based on its features. |

# **Exploratory Data Analysis**

Our main goal in this project is to build a model to predict sales price. Before we start variable selection, we need to explore our data to gain a greater understanding of how variables interact with one another. We will start by looking at the target variable ‘Sale Price’. It seems that our data set has a mean sale price of $186,761.78 with a Standard Deviation of $78,913.84. The plot below shows that the distribution of sales price is skewed to the right. We can see that most sales prices lie between $100,000 and $250,000 in the data set.

*Distribution of Sale Price*

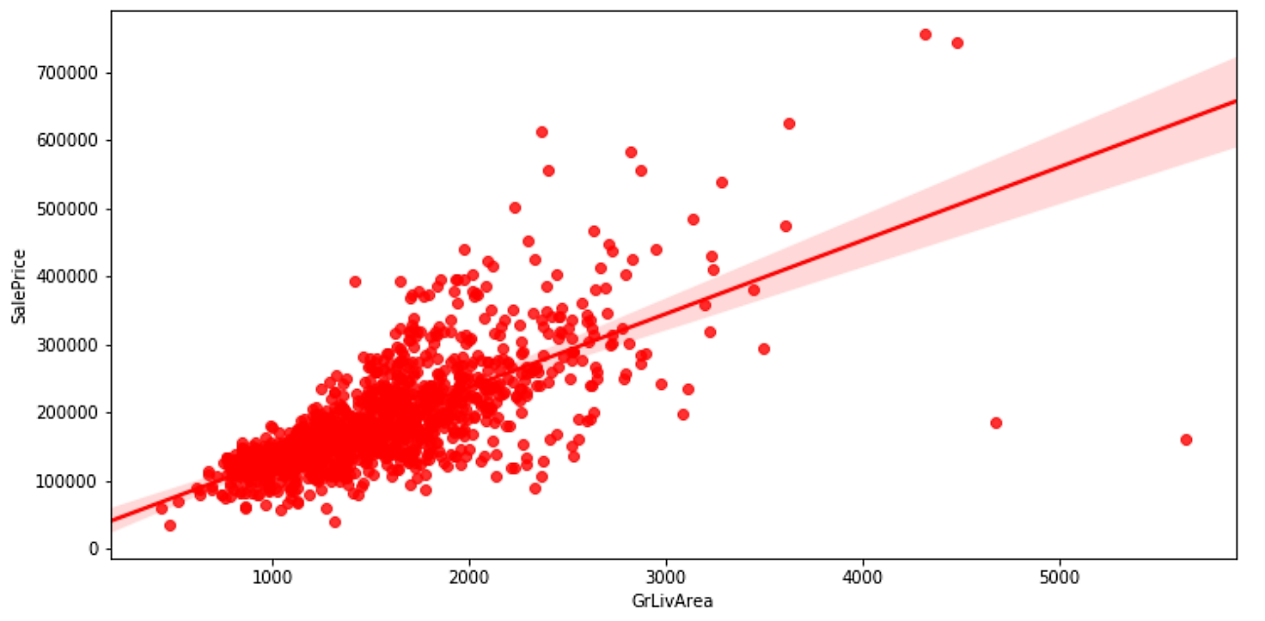


The First relationship I wanted to explore in the data set was the relationship between neighborhood and Sale Price. My initial thought is that there should be a strong relationship between neighborhood and Sale Price.

*Sale by Neighborhood*

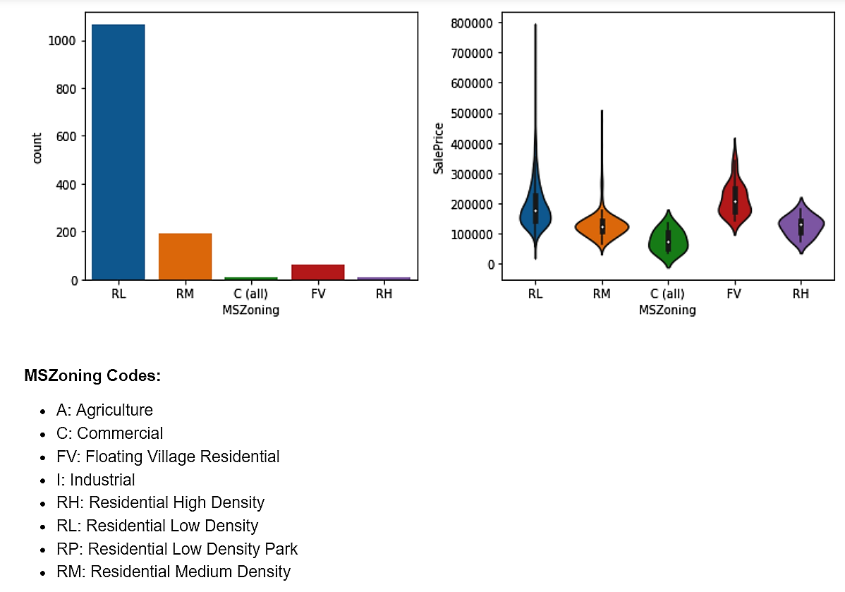
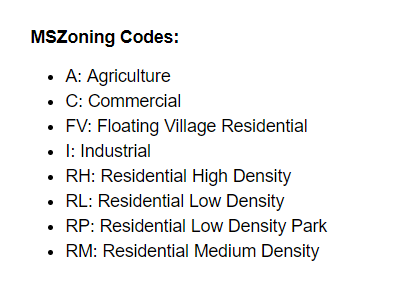
The ‘Sales Price by Neighborhood’ Plot shows ‘NoRidge’, ‘StoneBr’, and ‘NridgeHt’ are the neighborhoods with the highest average sales price. It seems that ‘BrDale’, ‘MeadowV’, and ‘IDOTRR’ are neighborhoods with the lowest average sales price. Upon further investigation of homes in these neighborhoods, it seems like these homes have lower than average ‘GrLivArea’ (Above ground living area in square feet) which could have contributed to the lower sale price of homes in this neighborhood. Looking at the Plot of GrLivArea and Sale Price, the plot reveals a high correlation between these 2 variables. The strong positive relationship leads me to believe that GrLivArea may play a crucial role in prediction model.

*Sale Price vs GrLivArea*



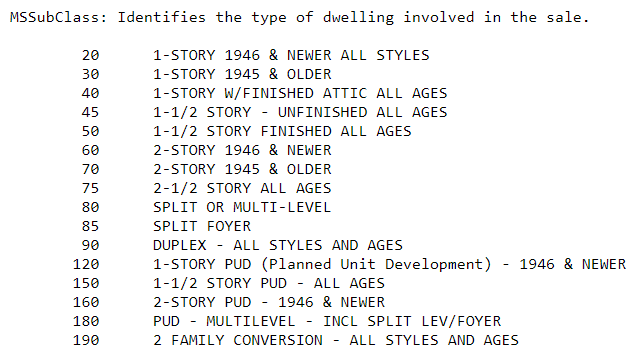
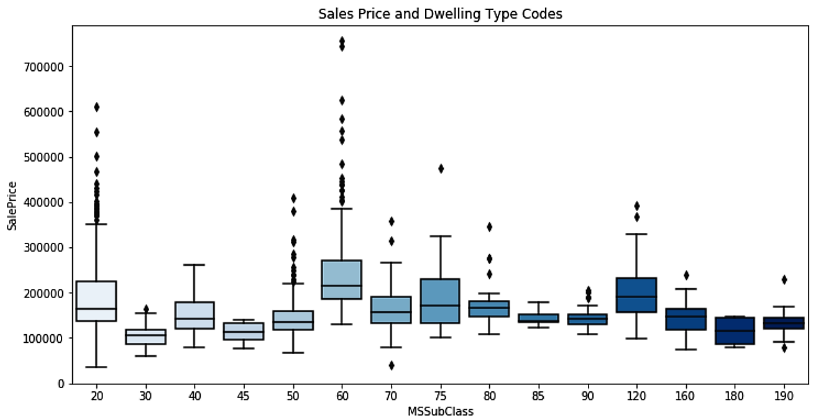
Moving on, I wanted to see if MS Zoning codes played a part in Sale Price. The chart below displays the MS Zoning codes present in the data set. The plot shows that most of our observations are homes that are in Residential Low-Density Areas. Homes with the MS Zoning code of ‘RL’ have an average sale price just under 200k and have the greatest variance of all MS Zoning codes. Floating Village homes are priced higher on average than any other code.

*Sale Price vs MS Zoning code*



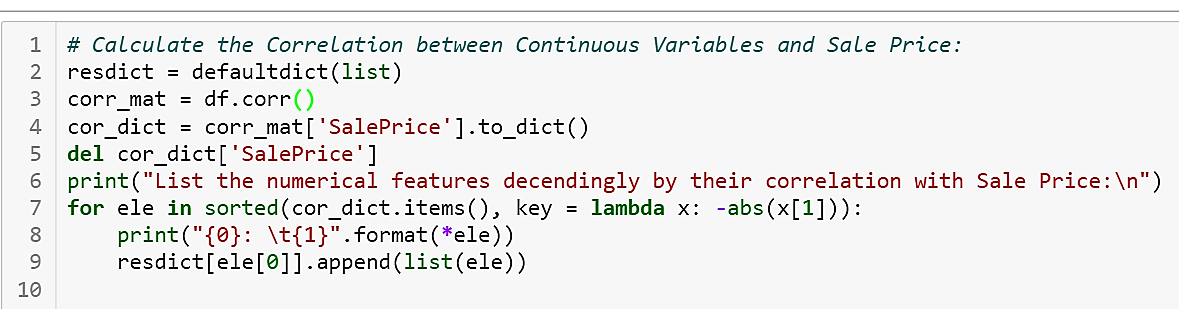
The chart below displays the sales price by dwelling code. The plot displays multiple outlier for each subclass. Notice that the code 60, which represents 2 story homes 1946 & newer, has the highest average Sale Price among all other dwelling code. The chart shows almost all dwelling codes have extreme outliers. Split Foyer homes (85) have the lowest variance in price when compared to other subclass types.

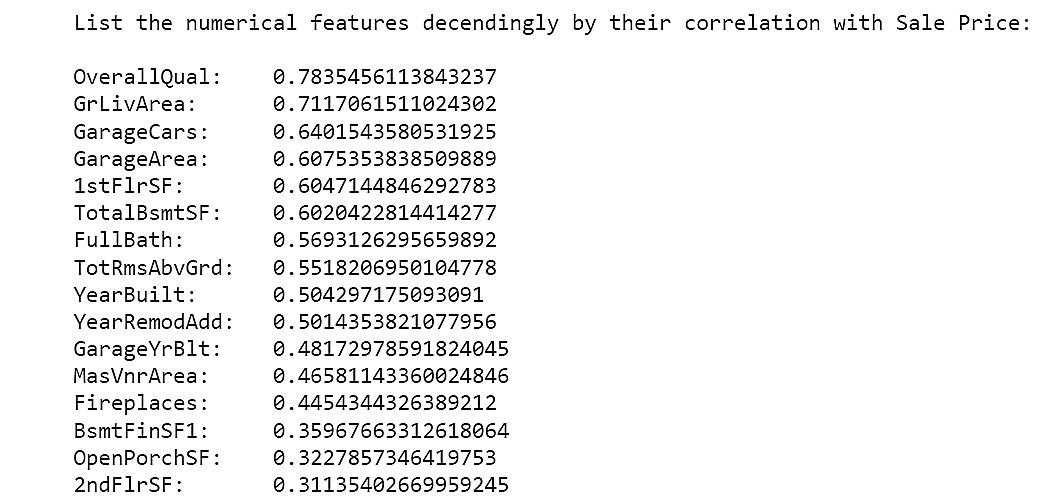
*Sale Price by MS Subclass codes*



# **Statistical Inference**

Our data set contains over 70 different predictor variables that can be used to predict sales price. Before the model is trained, we need to verify that all these variables have a statistically significant relationship with Sale Price. We start by looking at the continuous variables in the data set and looking at their relationship with sales price. I wrote a function that went through each continuous variable in the data set and calculates the correlation between that variable and ‘Sale Price’.



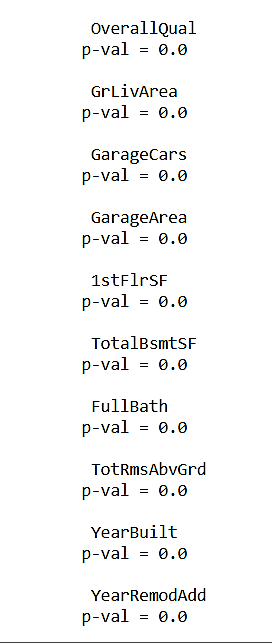
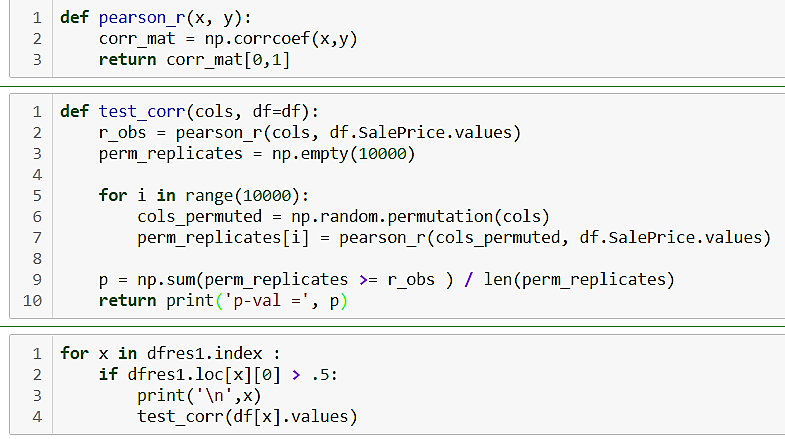


Overall Quality, GrLivArea, and Garage Cars have the strongest correlation with Sale Price. Before we include these variables in our model, we need to verify the validity of this relationship with a correlation test. The correlation test will verify that the relationship between Sale Price and a numerical variable isn’t by chance. The null and alternative hypothesis is as follows:

H0: There is no correlation between Sale Price and tested numerical variable

HA: There is a correlation between Sale Price and tested numerical variable

I used the following code to write a function that finds numerical features that have a correlation with Sale Price Higher than .50. Then I used a self-defined correlation function to iterate through the column names and test each one then lists the resulting p value.

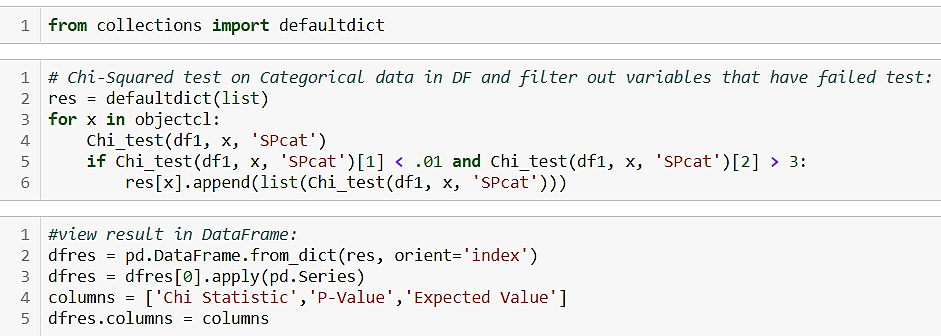


The results from our hypothesis testing shows that the above variables’ correlation with sales price are true and not by chance. We can be confident that these feature variables will play a crucial part in our model’s ability to predict Sale Price. These variables will be included in our initial model.

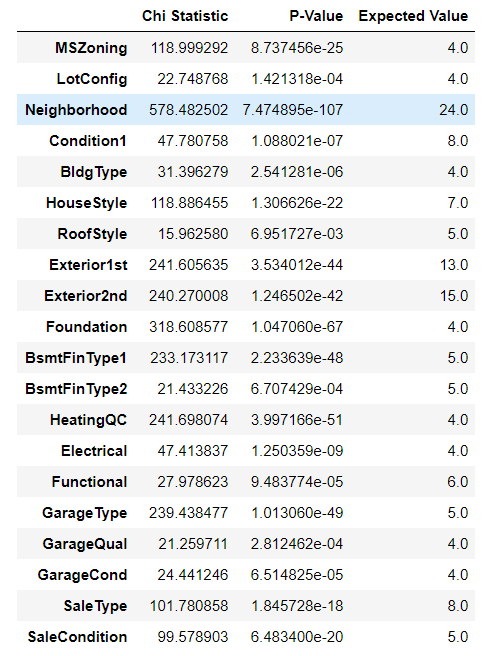
Moving on to the categorical variables, we will need to test the relationship with Sale Price as well. We can test this relationship using a chi-squared test of independence. To use this test, we need to categorize our sale price. Using the code below, I split Sale Price into 2 distinct groups. The Null and Alternate hypothesis for our testing is as follows:

H0: In the population, the tested categorical variables and categorized sale price are independent.

HA: In the population, the tested categorical variables and categorized sale price are dependent.



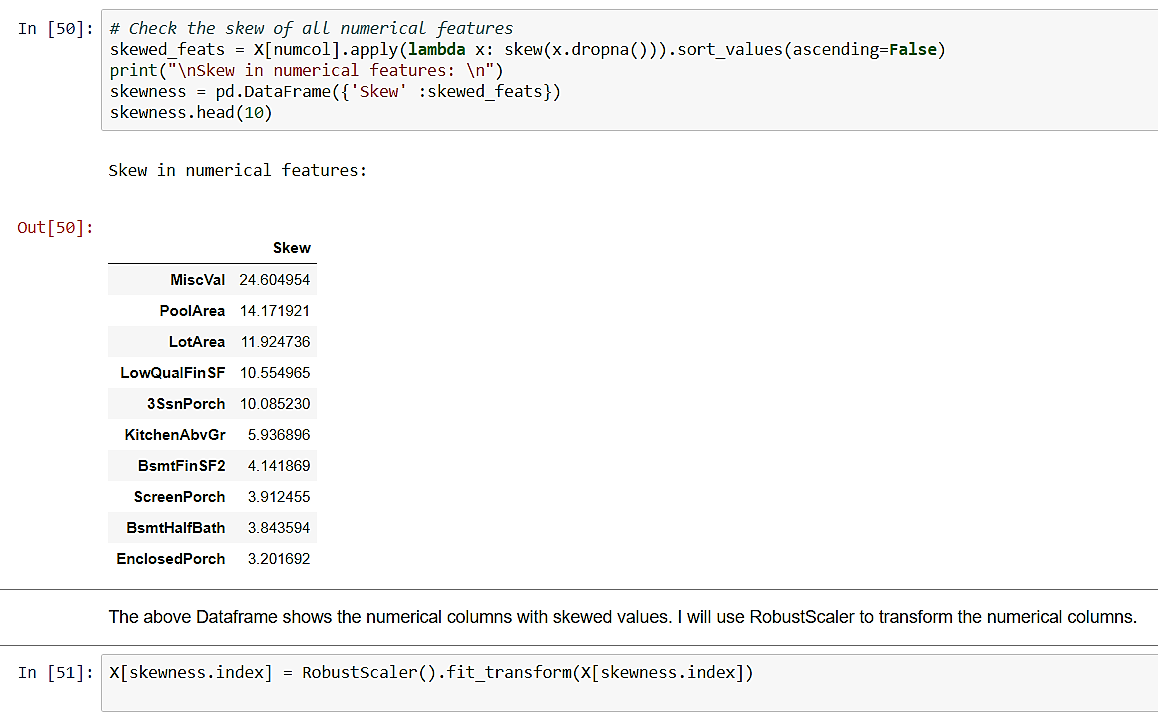
Next, I conducted the chi squared test of independence on each variable using the code below and append the results to a data frame that can be seen on the next page.



In conclusion, we have identified the variables that have a statistically significant relationship with Sale Price. From our analysis, the variables such as ‘GrLivArea’ and ‘Neighborhood’ share a strong relationship with sale price. We will use the variables we identified earlier in the report to develop a model that can best explain the variation is sale price.

# **Preprocessing and Model Construction**

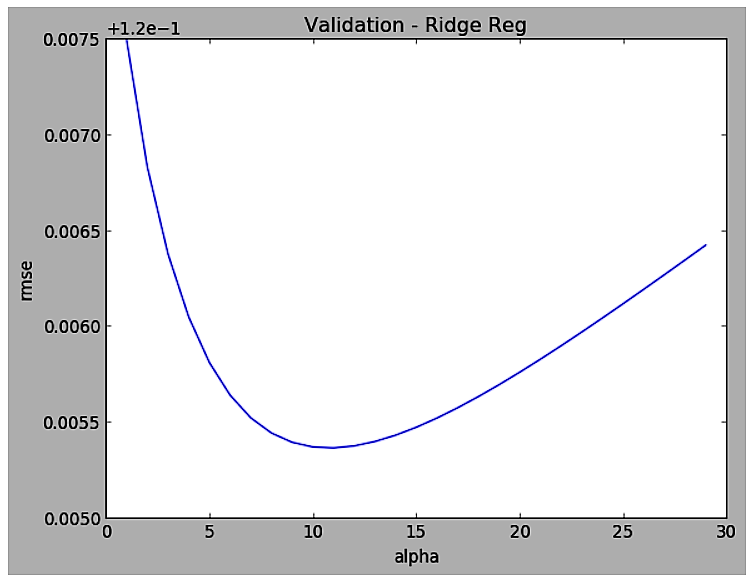
Now that the statistically significant predictor variables have been identified, The data needs to be prepped for model construction. We start by separating our predictors and target variables. Next, we identify skewed numerical features within the dataset. we need to scale the skewed numerical features to improve the model prediction power. Using the code below, I was able to identify the skewed features in the data frame and use function called ‘RobustScaler’ to scale feature variables based on their median to minimize influence of outliers. With numerical features scaled, I generate dummy variables for the categorical feature variables in the dataset. Then split the dataset into a training and test sets.



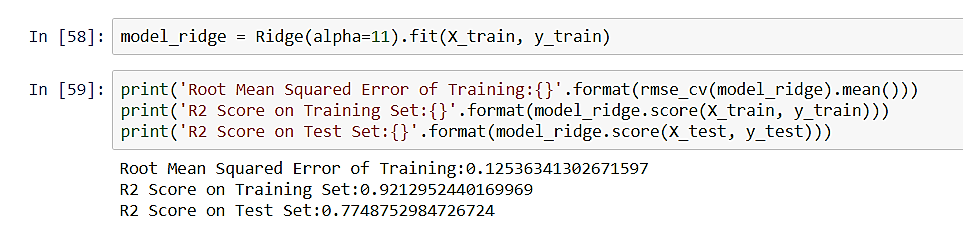
## 1st Model

In my first attempt, I tried to use Ridge Regression algorithm. This algorithm tries to retain all the features but shrinks coefficients. Since I have identified the variables to be statistically significant, I’m comfortable using a model of this complexity (many features). Before we train the model, I wrote a function that will evaluate the model using root mean squared error. This will provide an additional metric to compare different models. I began model construction by optimizing the hyper-parameter, The graph below shows the change in mean squared error across different alphas.

Ridge regression validation curve

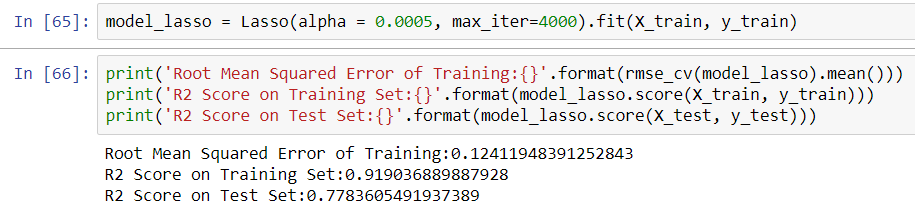


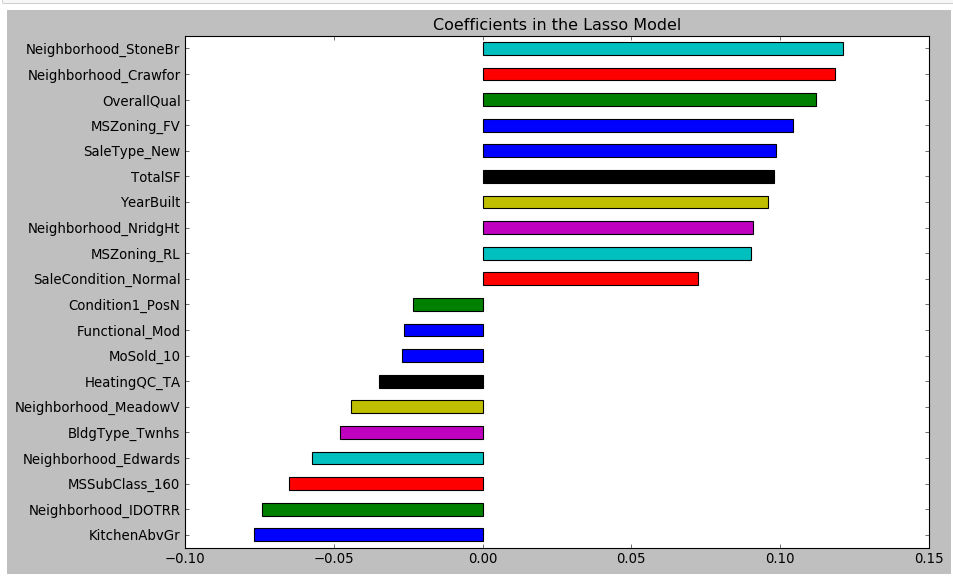
After I have identified the best parameter, I fit the model and used it to predict on the training and test set. The measure metrics listed below, shows that our model suffers from some overfitting. We will need to try and use a different algorithm that can provide more feature flexibility.



## 2nd Model

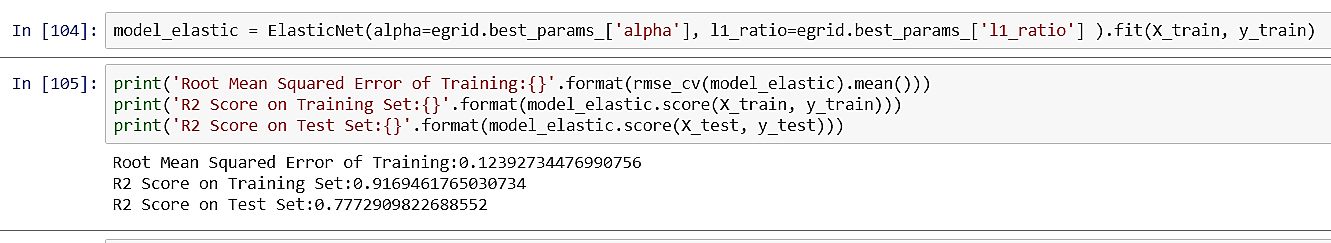
Moving away from Ridge Regression, I wanted to try a lasso regression algorithm to predict sale price. While fitting the model, Lasso regression algorithm automatically does coefficient reduction and feature selection. This algorithm should offer better predictions since it will use a subset of Key features to make prediction. As done previously, I optimized the hyper-parameters for the lasso algorithm and then fit the model to the training data. This model offered a slightly worse score on the training set but better prediction on the test set. Also, the root mean squared error decreased as well. The chart below shows the top feature selected for predictions.



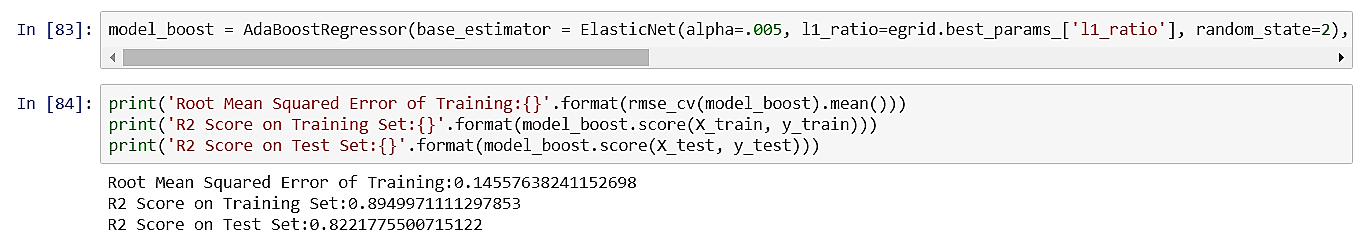
Feature Coefficient of Lasso Model

## 3rd Model and Boosting

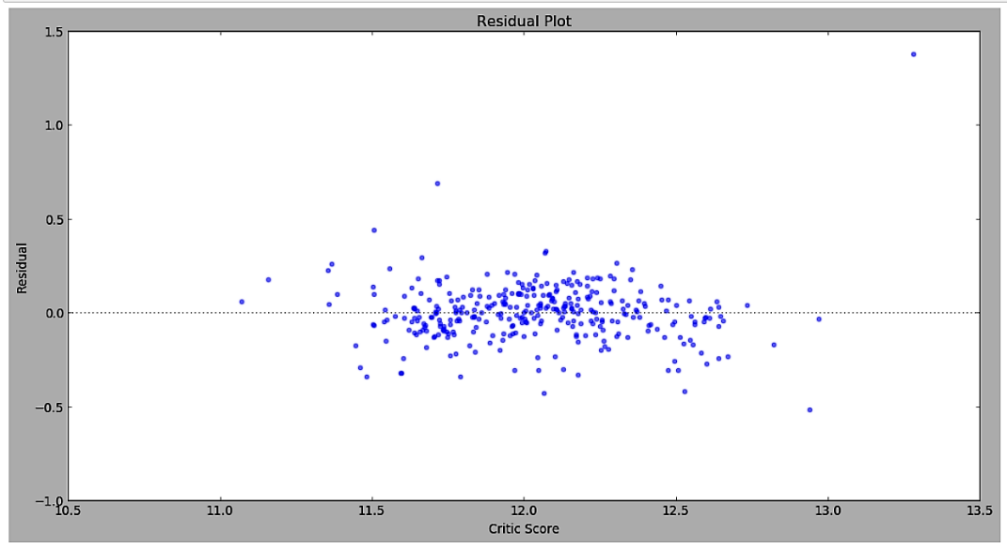
The lasso model performed better than the ridge model on the test set, but I believe that we could increase the explained variation by using a Hybrid of both model. Elastic Net Algorithm can provide that hybrid model. I start by optimizing the parameters then fit the model to the training data. This time the prediction on the test set slightly decreased and our error decreased as well.



At this point, our models power to predict on the test set hasn’t significantly increased. In one last attempt to improve upon the last model, I attempted to use the Ada boost algorithm to increase the prediction power of the model. The Ada boost algorithm fits a sequence of weak learners on different weighted training data. It starts by predicting original data set and gives equal weight to each observation. If prediction is incorrect using the first learner, then it gives higher weight to observation which have been predicted incorrectly. Being an iterative process, it continues to add learner until a predefined limit is reached in the number of models or accuracy. I will use Elastic net as my base estimator with slightly different parameters. I optimize the parameters for Ada boosting algorithm and fit using the training data. My last attempt was a success. The boosting algorithm resulted in a higher R-squared on the test set as seen below.



The residual plot below gives a visual representation of how well our model has performed. Notice the residuals a scattered somewhat evenly around the horizontal zero line, but there is an extreme outlier. Despite this outlier, I was still able to construct an effective model.

Residuals

# **Key Insights**

1. **Know the Squared Footage**
   1. Our engineered variable TotalSF turned out to be one of the strongest predictor of Sale price. The Sum squared footage of all floor in the home help determine Sale Price. Naturally the more space you have in the home the higher the sale price will be.
2. **Building Type Matters**
   1. Homes with classified as Townhouse Inside Unit tend to have lower sales price. These homes tend to have low total square footage. Avoid 2-STORY PUD - 1946 & NEWER homes since they tend to have a lower sale price. These homes seem to be the least desirable.
3. **Location, Location, Location** 
   1. Some of the Strongest predictors in the tied to the location of the home. Be mindful of the neighborhood and zone your home is located in. Depending on what neighborhood or zone your home is located, Sale price is greatly influenced. Within the data set homes that lied within the neighborhood of North Ridge and North Ridge Height had some of the highest sale prices on average. Also homes that reside within a Residential low-density zone had high sale prices as well.