House Prices: Advanced Regression Technique Final Report

Aakash Shah

I. Introduction

In this House Prices Advanced Regression Technique final report, we will be looking into the various parameters of House Prices and predict the scores which will evaluate whether the house prices will be accurate to the actual price. Furthermore, we will create and evaluate six various regression models (Random Forest and Gradient Boosting, etc.) Our Project Goal is to evaluate which of the six models created would give us a best score to predict the house sale prices.

A) Criteria for success

"Submissions are evaluated on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)" - Kaggle

B) Stakeholders

Project stakeholders include the following: home buyer's, landlords, the Ames Iowa Village Officials, Realtors, and surrounding residents.

C) Data sources

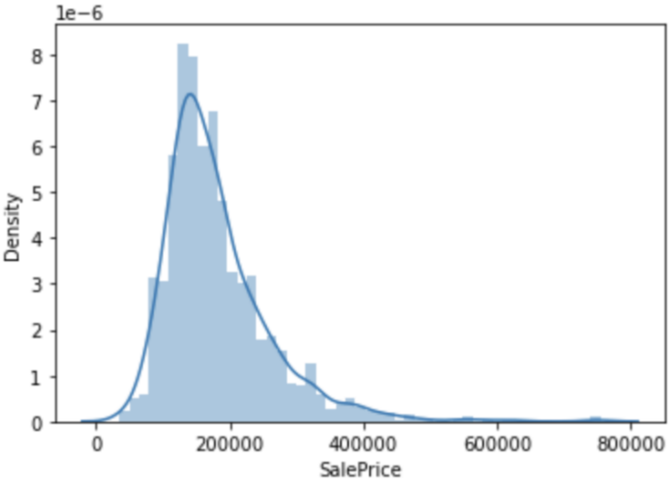
Kaggle- Ames Housing Dataset compiled by Dean Se Cock

D) What will your client do or decide based on your analysis?

Our clients here are the home buyers, who care about the problem to understand the lowest house prices to understand to see if it fits their budget. All of our stakeholders can use it for various use such as the following:

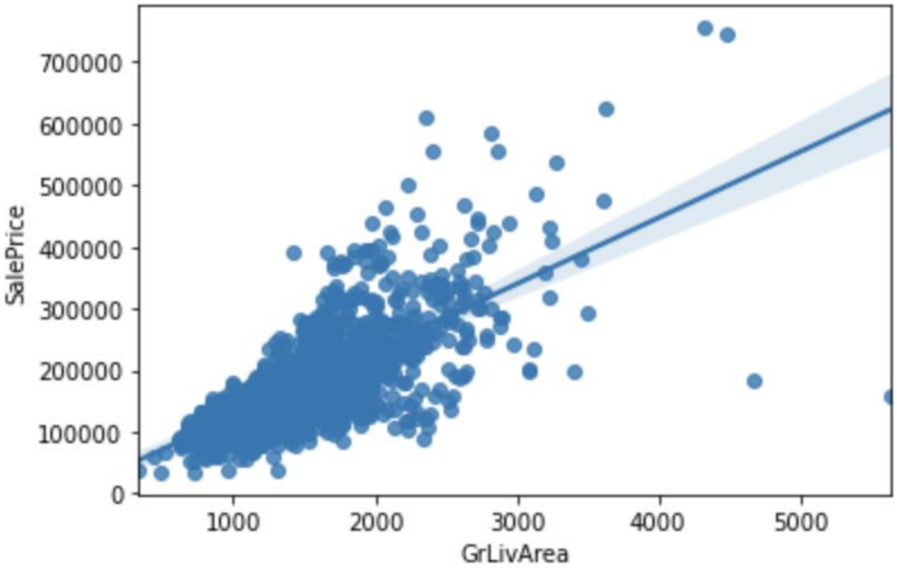
2) Exploratory Data Analysis

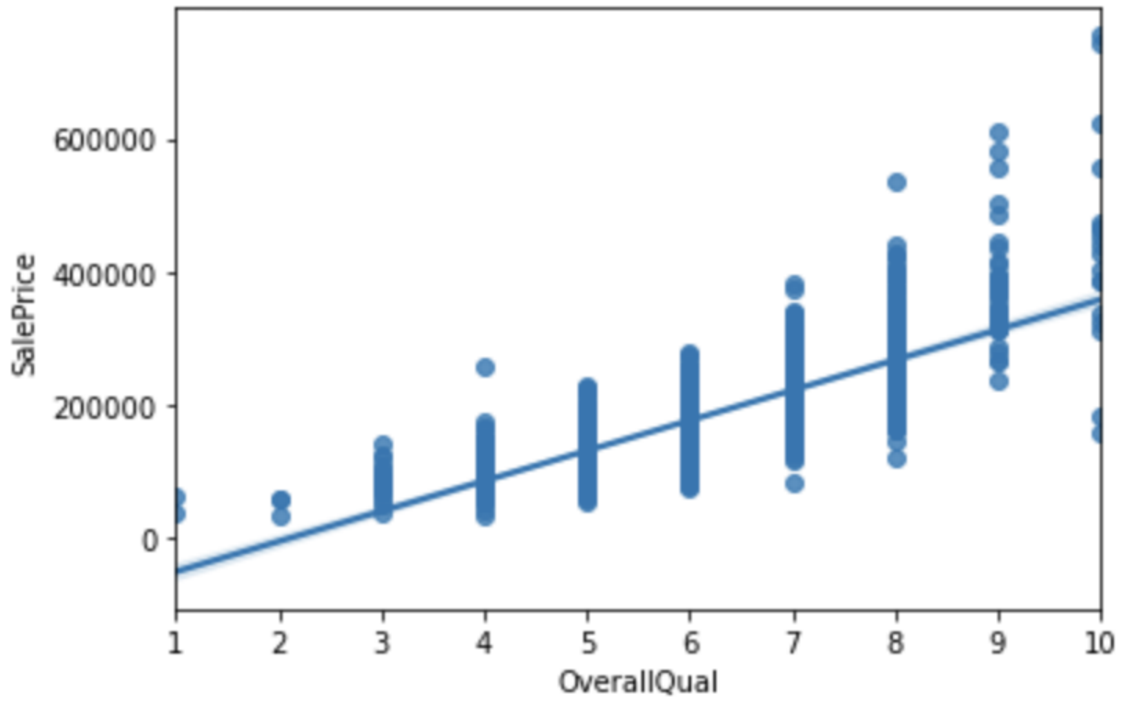
This Exploratory Data Analysis was conducted earlier in the process to identify which features may be more important than others. Seen on the left- we first found the distribution of the target variable: SalePrice, to understand where the chart lays in terms of skewness and to see if the data is normally distributed. In this case, we see that the graph is right tailed and that we would have to adjust that to fix that to make sure that nothing is going to mess with the data which has been provided to us. Once that was completed, we had selected the categorical features and created the heatmap to see the outliers of the features.

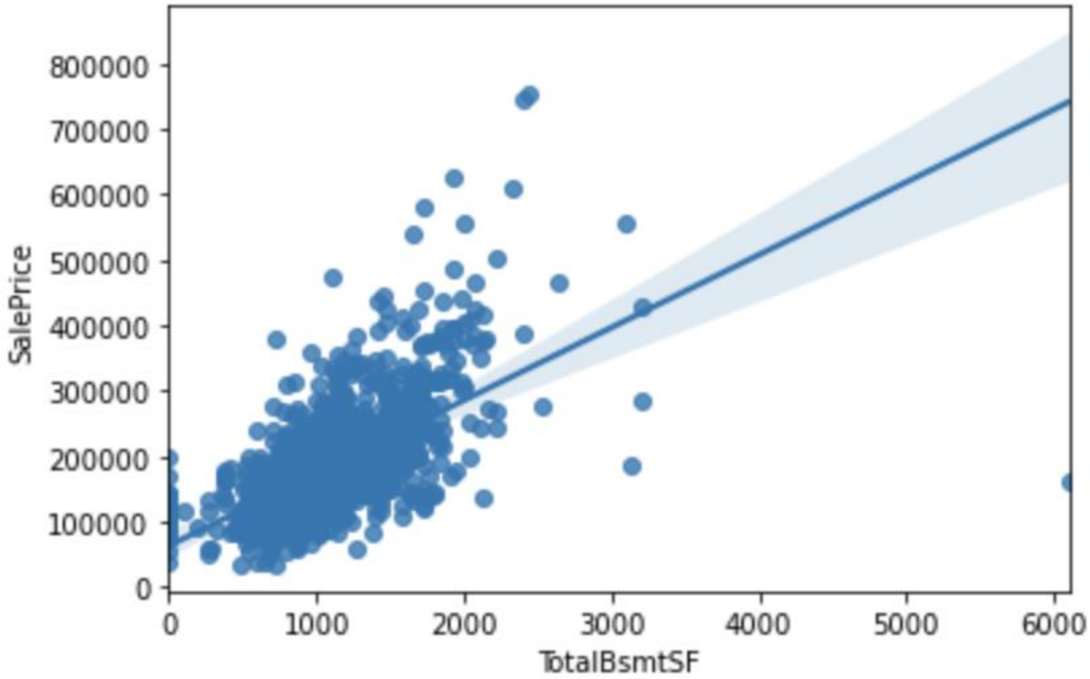


A) Scatter Plots

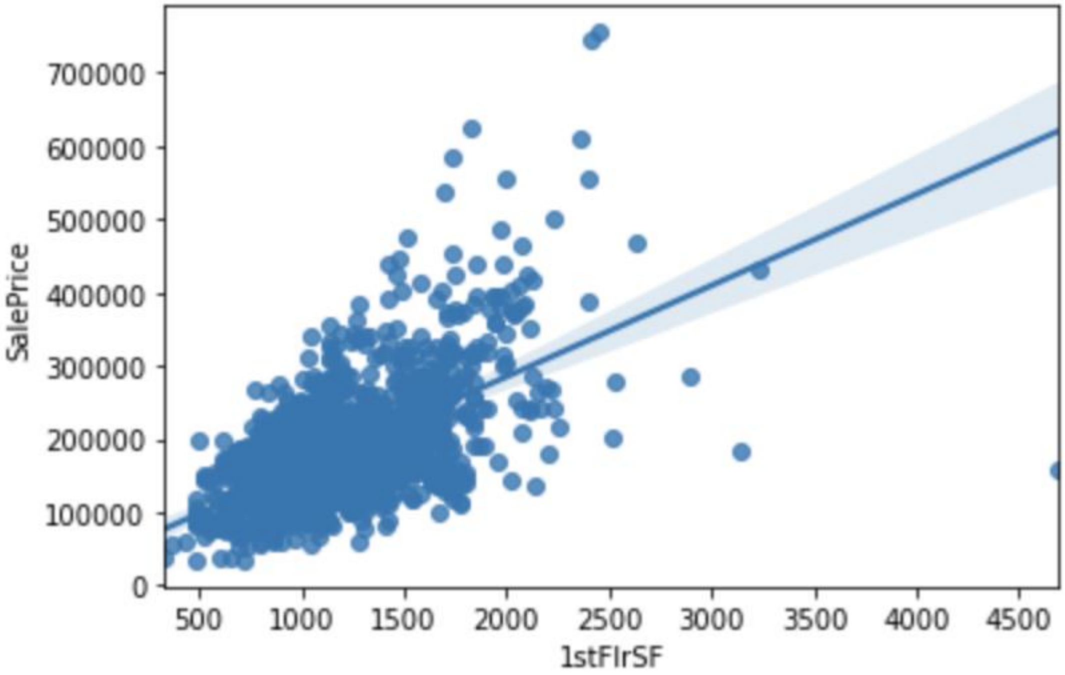
When comparing the GrLivArea against the SalePrice, we see the outliers of a higher living square area at a lower SalesPrice. Through the illustration we see that the average prices sit at $300,000 with 2000 square feet living area. The second image shows A simpler correlation which shows us that the higher the SalesPrice had gone up, the higher the OverallQuallity per house. The third image shows us that the Total square feet of basement area against the SalePrice; positive correlation of the more Total square feet of basement area, the more value and price which the owner would be paying for.

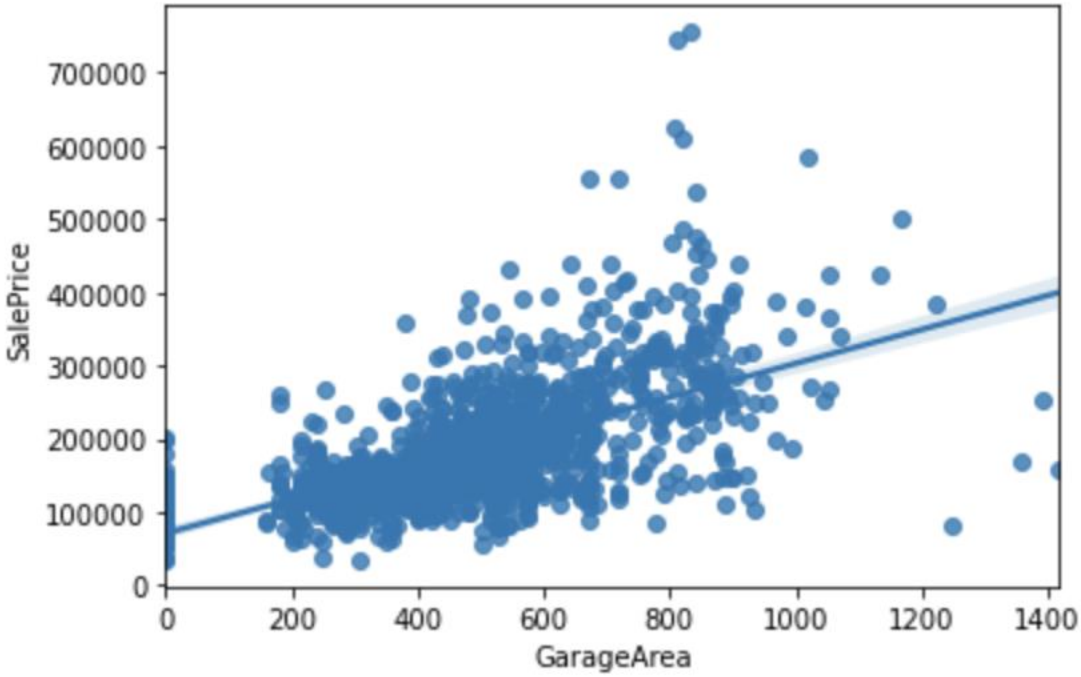



SalePrice
600000
400000
200000
0
1
2
3
4
5
6
7
8
9
10
OverallQual

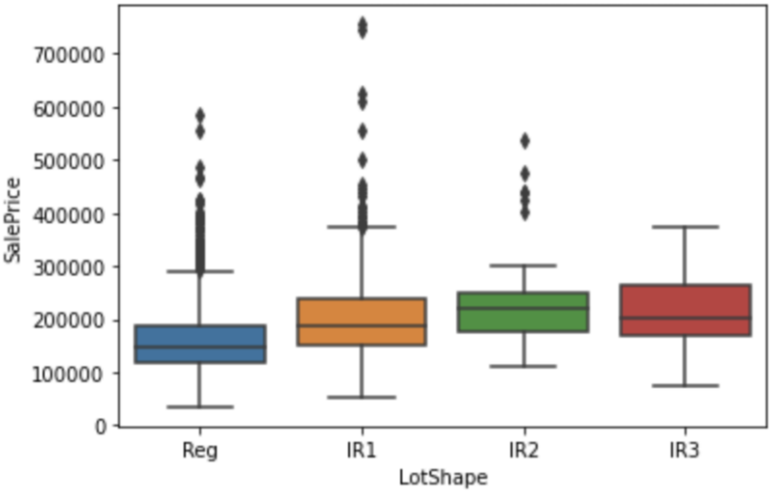

SalePrice
800000
700000
600000
500000
400000
300000
200000
100000
0
0
1000
2000
3000
4000
5000
6000
TotalBsmtSF

Shown in the scatter plots below, below, we see a similar trend and pattern going on with the SalePrice and the specific feature, as the SalePrice goes us, the feature goes up. For these next two scatter plots, we see that the outliers represent that those specific house type will be rare with that specific given feature.



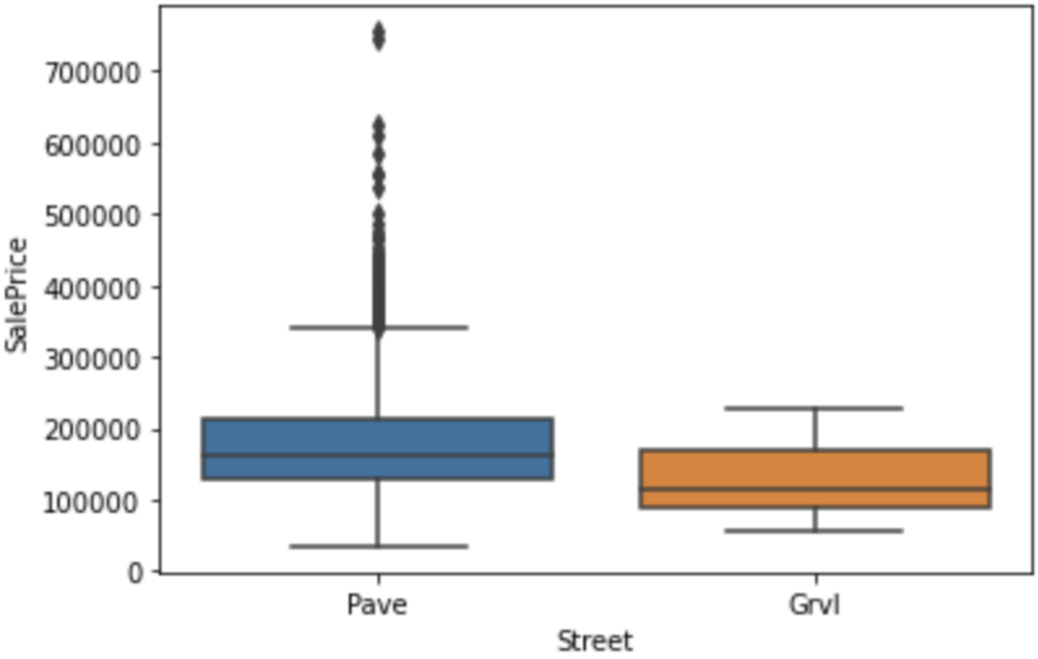


B) Box Plots

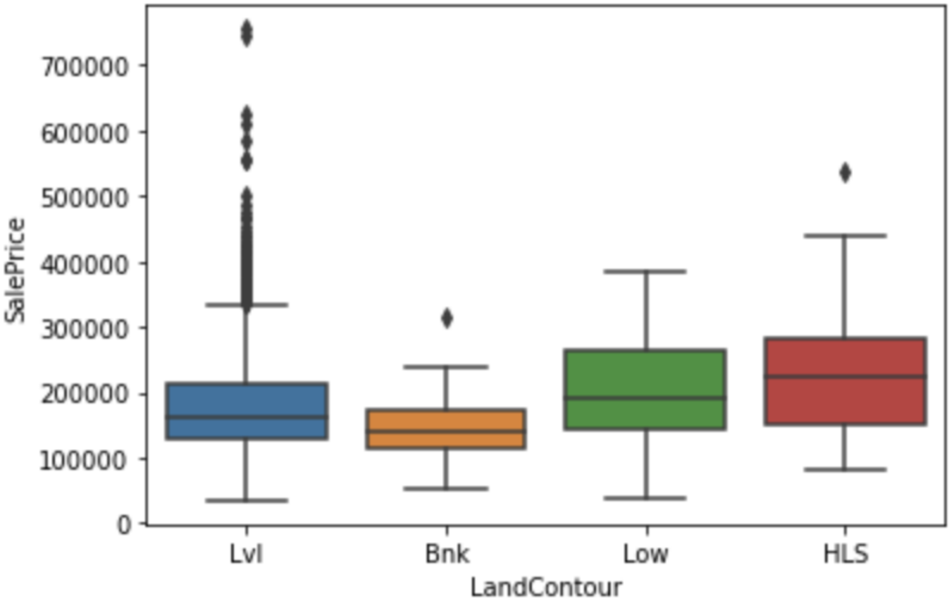


As any other box plot, we see the means that they lie in. In the boxplot to the left, the illustration states that the average of the Saleprice is very similar dependent on the four different lot shapes. Furthermore, we see that the formality of the data looks normal considering that the mean forms the boxplots are close to each other. According to this diagram, we see that IR2 has the highest mean out of the three. This considers that IR2 may have more features to the lot, which is why the mean may sit higher than the rest.

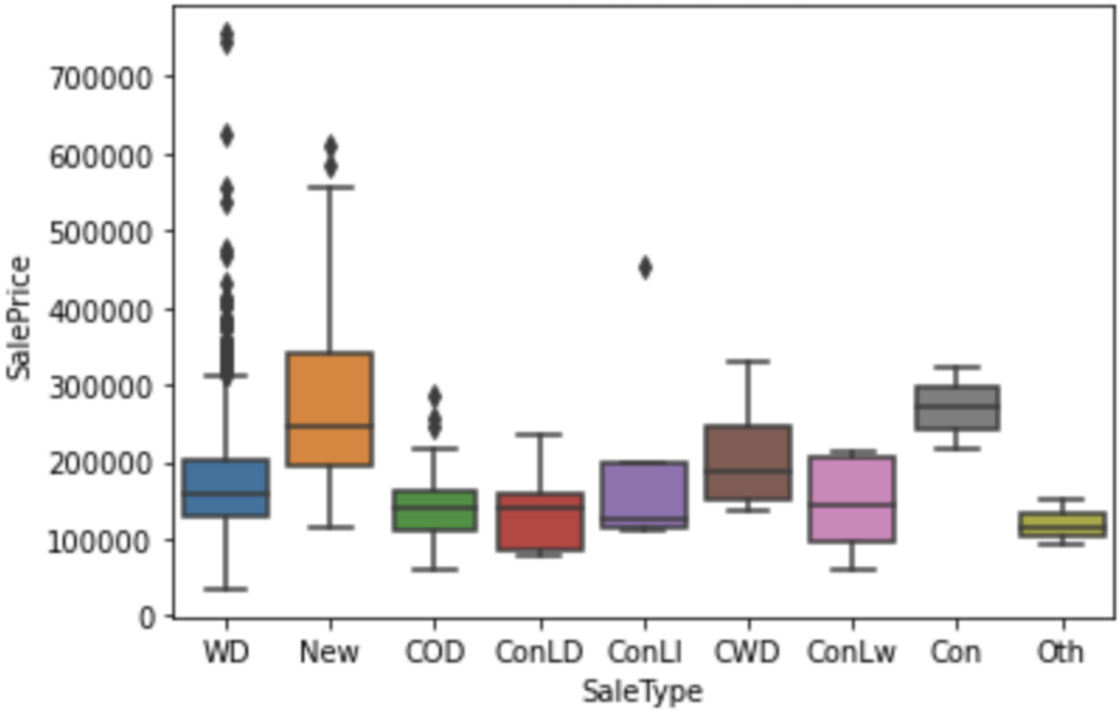
As we have a similar case to the boxplot below, we see that the mean is close to each other and that the skewness in the two plots below are pretty positive, as there are no outliers. Lastly, here we see that the skewness is symmetrical.



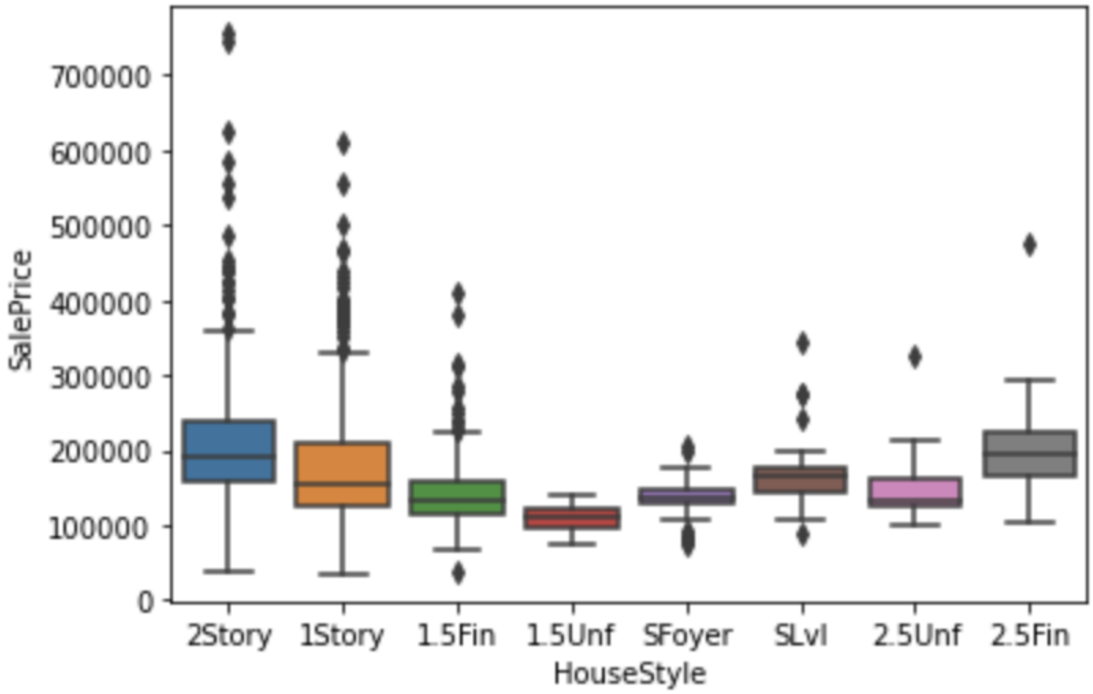
For the two box plots below, similarly to. Our typical trend of boxplots, our HL5 looks particularly high but the means are close together.

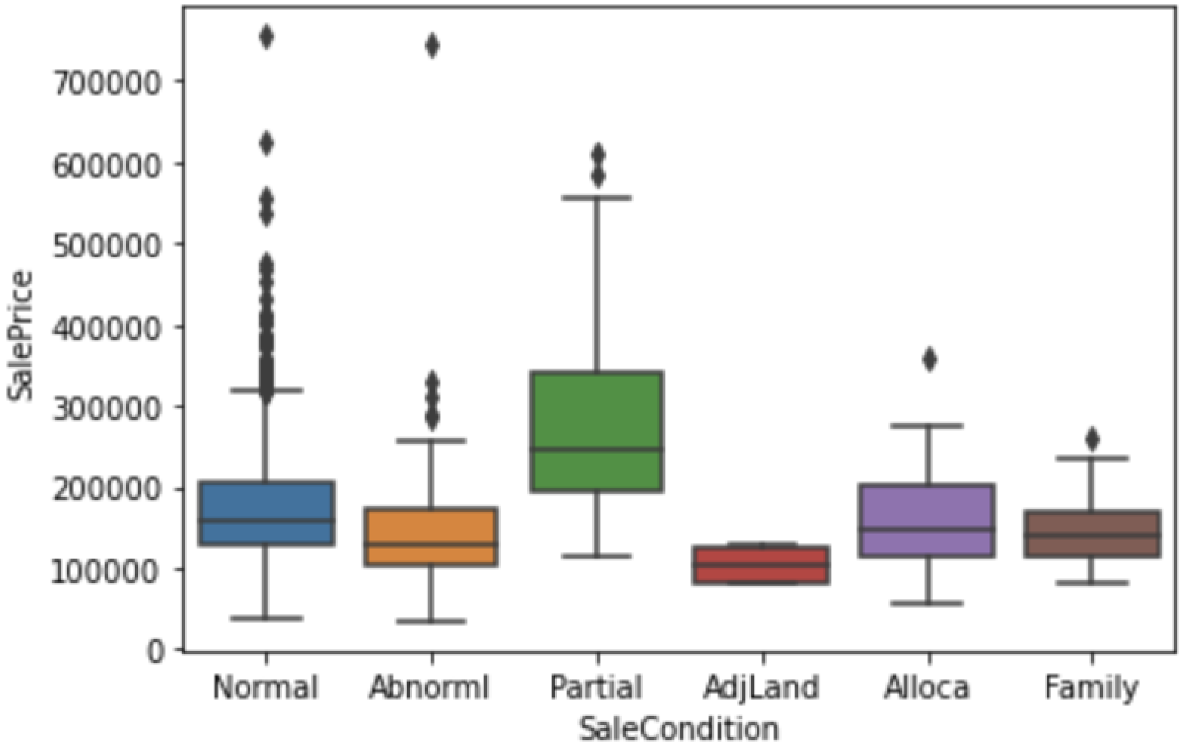


For our SaleType's there are many various medians. The closer ones which one may choose are the COD, ConLD, ConLI, and CWD. As many as the others are fairly high in the mean, those are more of the odd outliers which still may remain.



In the two chars below, same but similar distributions in our HouseStyle feature, with a couple high means in the SaleCondition. Illustrations show the correlated SalePrice against each specific feature. Something to maybe keep in mind is that fi there is a high mean, there can be less people interested.





3) Data Wrangling

For this Data wrangling, we have created a Pandas HTML overview report, which has illustrated the general overview, warnings, and the reproduction for some of the models. In the overview, it gives us all the key information such as number of variables, number of observations, missing cells (in which case was 0 in the dataset), number of duplicate rows, and the memory size. The Pandas Profiling Report also has shown us the four types of correlations of heat maps to have a better understanding of the correlations and outliers.

To better clean up the data, we change and looked for all the numeric data in the Train data set. This was done by selecting float64, and int64. We then had dropped the duplicates and found the correlation of the largest value of the target. The rest of the categorical values which we did not clean up in the stage, was a part of the process in the preprocessing stage.

4) Preprocessing & Modeling

In the preprocessing stage, I had converted the data frame into dummy variables so that I can include all the categorical values in the data frame. I then split and trained the data for the various models. In the modeling stage, I had created a Linear Regression, L2: Ridge Model, L1: Lasso Model, Random Forest Regressor, a Decision Tree Regressor, and a Gradient Boosting Regressor. My focus is to create these models so that I can compare which of the six models have the lowest score. We take the lowest score because it provides us the highest accuracy of the prediction between the actual and the predicated data. The results in the data showed us that the Gradient Boosting was the best accuracy of the prediction and the actual of the predicated data. When building the models, I was also trying to only find the Mean Squared Errors.

5) Stakeholders

Over the course of the project, it can be further implemented by home buyer's, landlords, the Ames Iowa Village Officials, Realtors, and surrounding residents. Our analysis of the overall Data Science Method can be used to understand the SalePrice's of the dataset are fairly accurate to the actual houses. This would help the stakeholders make new features if there is another factor which is evolved in the housing market and or understand the impact that the price can make to the buyers.