Chad Warner

FINAL REPORT:
Predicting Housing Prices Using Machine Learning Techniques

TABLE OF CONTENTS

Problem Statement

The word “home” can take on a myriad of different meanings to different people. For the most part, home means an enjoyable, happy place where you can live, laugh, and learn. It’s somewhere where you are loved, respected, and cared for. When you look at it from the outside, home is just a house. A building. A structure. But on the inside, it’s a lot more than wood, bricks, and aluminum. https://meaningofhome.ca/past_winner/the-true-meaning-of-home. Determining the value of a house is important to both the sellers of houses as well as the buyers of

houses. Currently, there are several traditional ways in which the entities that control the housing markets determine housing prices. Some of these methods include, but are not limited to:

1) **Use the Federal Housing Financial Agency (FHFA) House Price Index Calculator -** The tool uses the "repeat sales method." Armed with millions of mortgage transactions gathered since the 1970s, the FHFA tracks a house's change in value from one sale to the next. Then it uses this information to estimate how values fluctuate in a given market.
https://www.nerdwallet.com/article/mortgages/how-to-determine-home-value.

2) **Obtain a Comparative Market Analysis (CMA)** - A comparative market analysis is a tool that real estate agents use to estimate the value of a specific property by evaluating similar ones that have recently sold in the same area. It can be extremely challenging to reliably estimate the fair market value of a home because there are a significant number of factors that go into determining how much a specific property is worth. https://www.rocketmortgage.com/learn/comparative-market-analysis.

3) **Hire a Professional Appraiser** - An appraisal is an unbiased professional opinion of the value of a home and is used whenever a mortgage is involved in the buying, refinancing, or selling of that property. A property's appraisal value is influenced by recent sales of similar properties and by current market trends. The home's amenities, the number of bedrooms and bathrooms, floor plan functionality, and square footage are also key factors in assessing the home's value. The appraiser must do a complete visual inspection of the interior and exterior and note any conditions that adversely affect the property's value, such as needed repairs.https://www.nerdwallet.com/article/mortgages/how-to-determine-home-value.

Although, there is no exact science to determine the value of housing prices, some of the aforementioned methods have positive impacts as well as negative impacts. However, these methods seem to be either subjective and arbitrary or not precise or accurate enough to apply to each home individually.

We are now in the digital age. With the advent of big data over the past two decades, more data is readily available and technology allows us to provide a more individualized detailed analysis in order for us to make more informed decisions. Machine learning algorithms have the ability to remove much of the arbitrary calculations and utilize metrics to predict housing prices based on certain features or attributes of different homes.

In this project, I was able to use a dataset, compiled by Dean De Cock, describing the sale of residential property in Ames, Iowa from 2006-2010. Based on the features contained in the dataset, I was able to create regression models that were able to predict housing prices in that market with minimal error. Using these machine learning techniques may have the ability to replace independent appraisers and bank property appraisers and save the homeowner as well as the individual money on not having to hire personnel with subjective opinions and calculations.

Data Wrangling

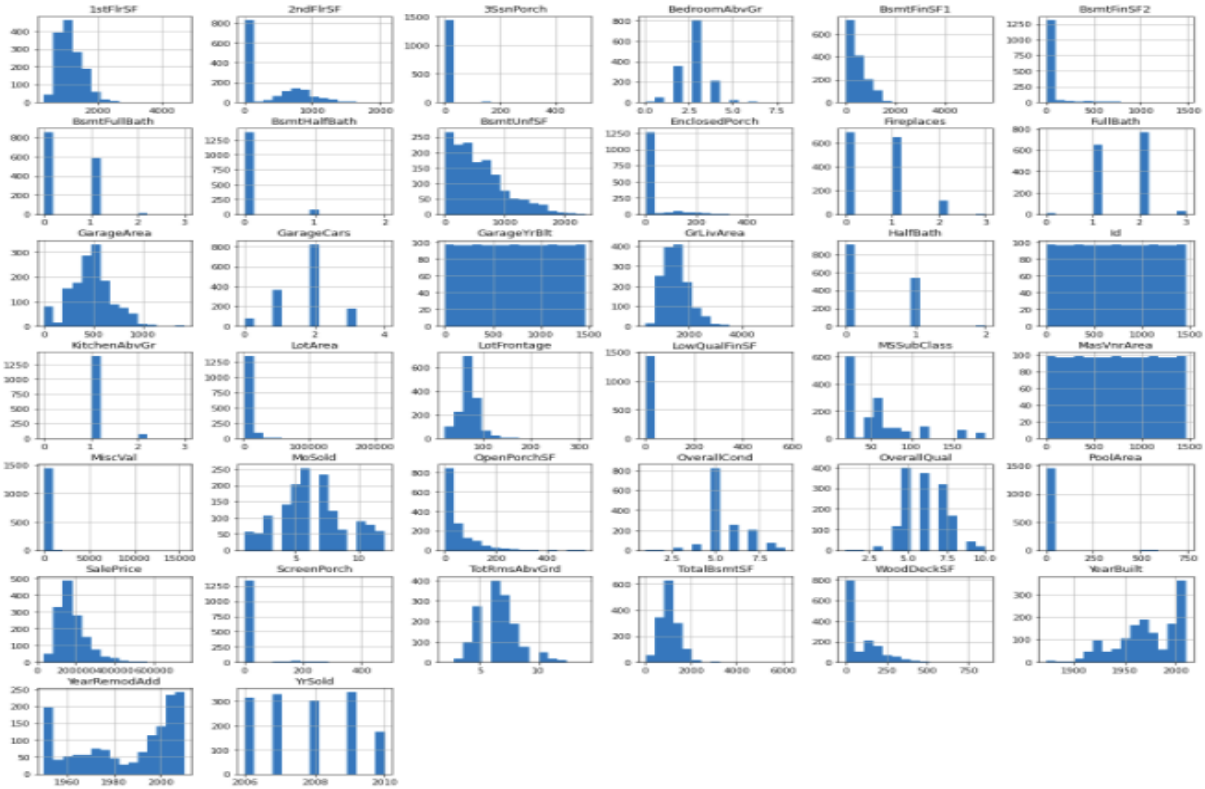
This complete dataset contained 2930 individual observations with a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involved in assessing home values. However, this dataset was split into a training set and a test set. I wrangled the data in the training set. So The structure of this dataset was a bit complex as all of the data was not in the same form. First, I checked the dataset for null values.

```
PoolQC          99.52
MiscFeature     96.30
Alley           93.77
Fence           80.75
FireplaceQu     47.26
LotFrontage     17.74
GarageYrBlt      5.55
GarageType       5.55
GarageFinish     5.55
GarageQual       5.55
GarageCond       5.55
BsmtFinType2     2.60
BsmtExposure     2.60
BsmtFinType1     2.53
BsmtCond         2.53
BsmtQual         2.53
MasVnrArea       0.55
MasVnrType       0.55
Electrical       0.07
dtype: float64
```

The above table illustrates the percentage of missing values in each column of the dataset. I then created a missingness matrix to create a visual on the amount of incomplete data in the dataset. I had the choice of whether or not to delete complete columns based on a large number of missing values or to impute the missing values in the column with a given technique. I chose to fill the missing values with the word "None" if a particular observation did not contain a certain type of feature. Next, I filled the missing values in the "Lot Frontage" column with the mean value of the "Lot Frontage" in the respective neighborhood. Finally, for the "Electrical" column, I imputed the missing value with the most frequent value contained in that column.

Exploratory Data Analysis

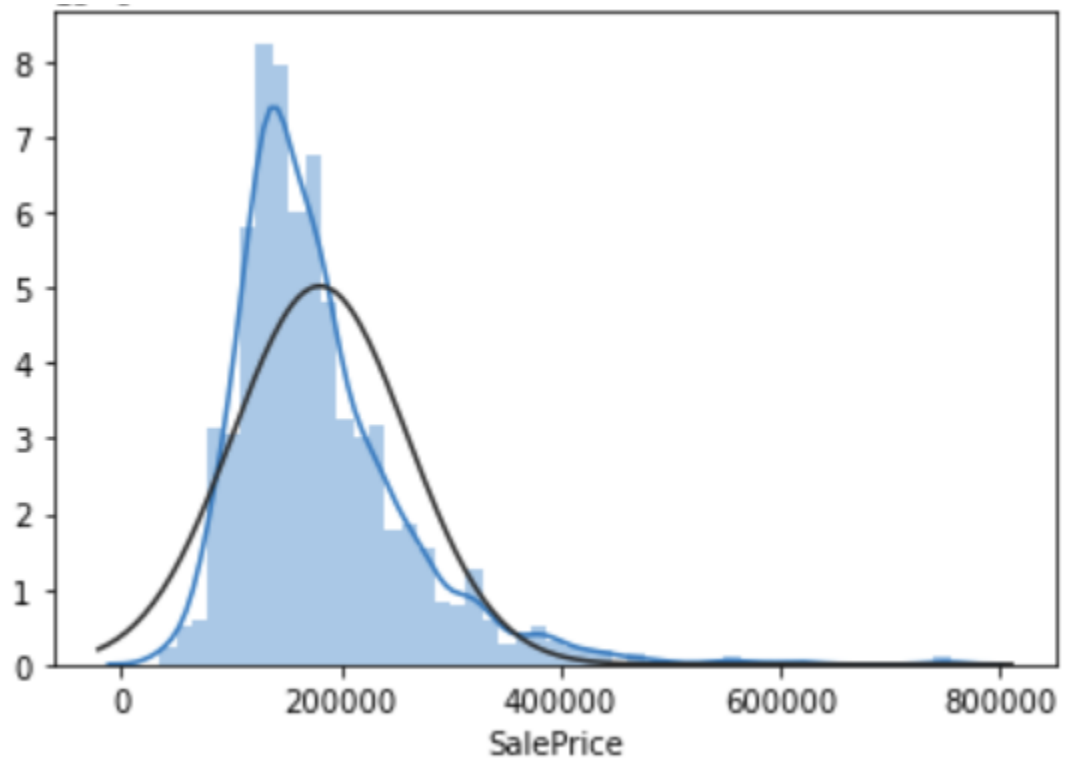
After I cleaned up the data and either deleted columns with a high number of missing values or imputed the missing values, I checked to make sure that all of the columns had observations relative to the observation. Roughly half of the entire dataset contained either integers or floating numbers and the other half of the dataset contained objects. In order to perform an informative exploratory data analysis, I decided to group the objects together and evaluate them and then to group the numerical values together and evaluate them. I created histograms of all of the numerical features of the dataset, in order to attempt to identify any relationships, which can be seen below:



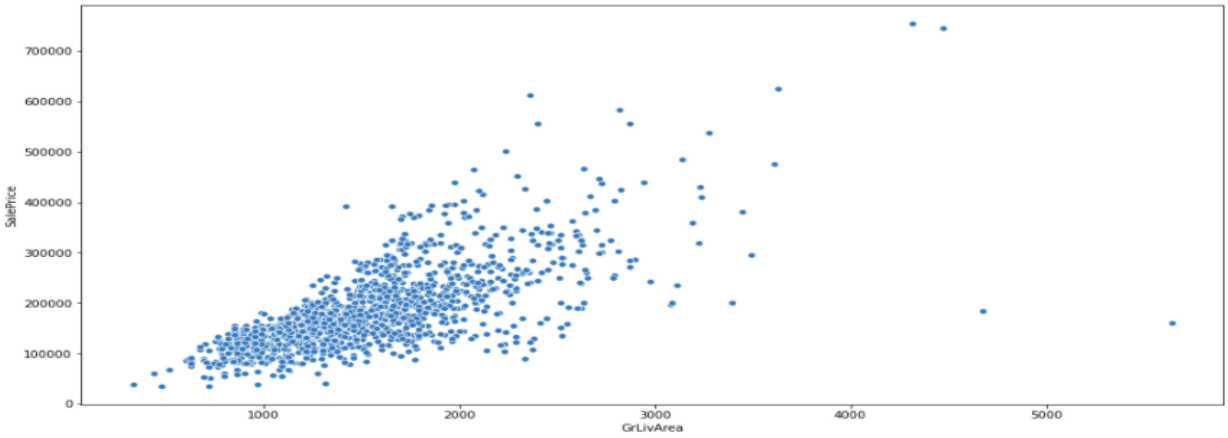
Afterwards, I did the same for the non-numerical objects, which can be seen below:


count
MSZoning
RL RM C (all) FV RH
Street
Pave Grvl
Alley
None Grvl Pave
LotShape
Reg IR1 IR2 IR3
LandContour
Lvl Bnk Low HLS
Utilities
AllPub NoSeWa
LotConfig
Inside FR2 Corner CulDSac FR3
LandSlope
Gtl Mod Sev
Neighborhood
Condition1
Condition2
BldgType
2fmCon Duplex TwnhsE Twnhs
HouseStyle
RoofStyle
Gable Hip Gambrel Mansard Flat Shed
RoofMatl
Exterior1st
Exterior2nd
MasVnrType
BrkFace None Stone BrkCmn
ExterQual
Gd TA Ex Fa
ExterCond
TA Gd Fa Po Ex
Foundation
PConc CBlock BrkTil Wood Slab Stone
BsmtQual
Gd TA Ex None Fa
BsmtCond
TA Gd None Fa Po
BsmtExposure
No Gd Mn Av None
BsmtFinType1
GLQ ALQ Unf Rec BLQ None LwQ
BsmtFinType2
Unf BLQ None ALQ Rec LwQ GLQ
Heating
GasA GasW Grav Wall OthW Floor
HeatingQC
Ex Gd TA Fa Po
CentralAir
Y N
Electrical
SBrkr FuseF FuseA FuseP Mix
KitchenQual
Gd TA Ex Fa
Functional
Typ Min1 Min2 Mod Maj1 Maj2 Sev
FireplaceQu
None TA Gd Fa Ex Po
GarageType
GarageFinish
RFn Unf Fin None
GarageQual
TA Fa Gd None Ex Po
GarageCond
TA Fa None Gd Po Ex
PavedDrive
Y N P
PoolQC
None Ex Fa Gd
Fence
None MnPrv GdWo GdPrv MnWw

Next, I evaluated the target variable (i.e. Sale Price). The minimum Sale Price in the dataset is $34,900. The maximum Sale Price in the dataset is $755,000. I then decided to plot the distribution of the target variable.

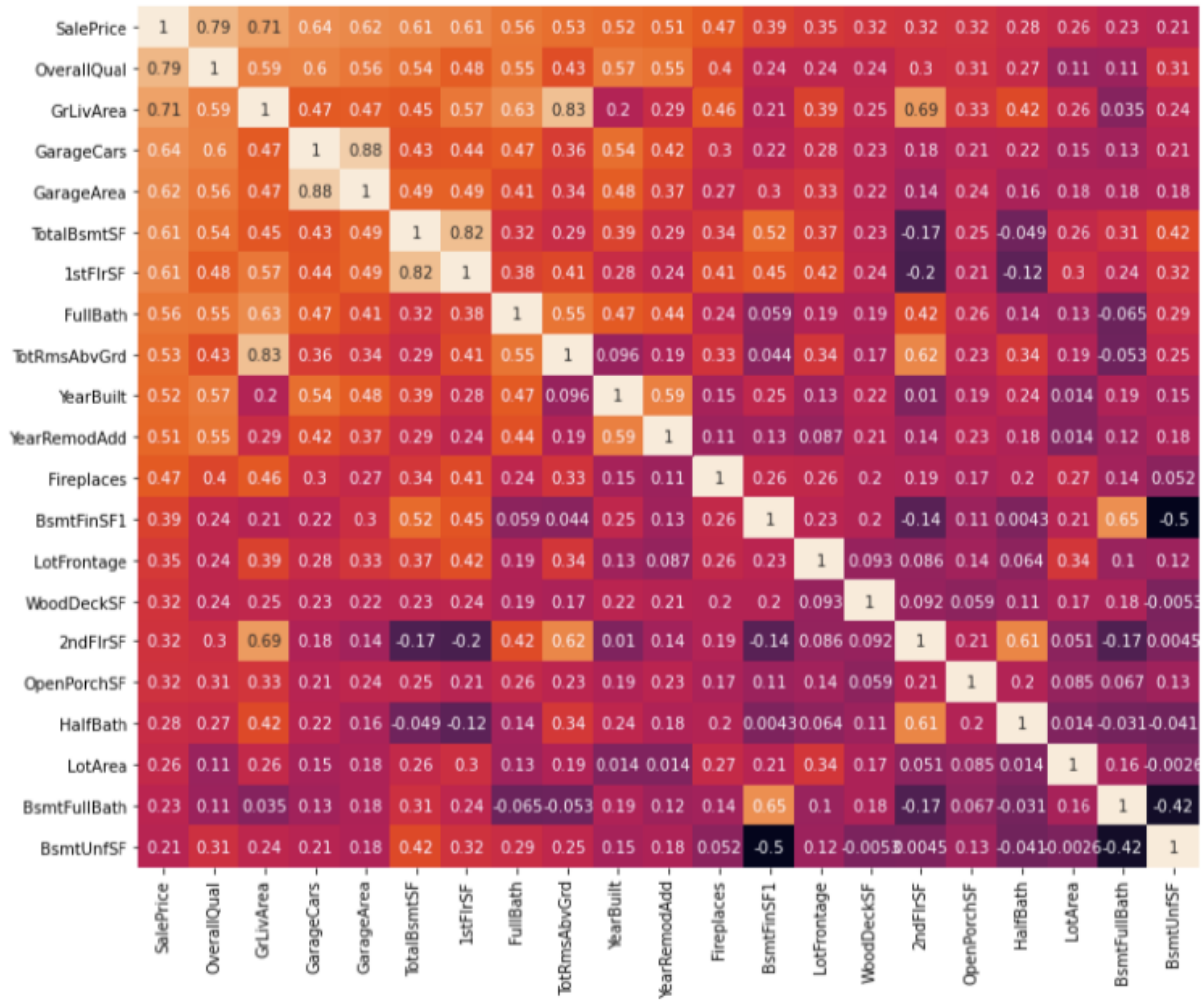


The distribution of the target variable (Sale Price) seems to be right tail skewed. An interpretation of this is that there are more samples of homes with lower sale prices than there are samples of homes with high sale prices. I then decided to plot some variables to see what type of relationship they had with the sale price. First, I plotted Greater Living Area against Sale Price.

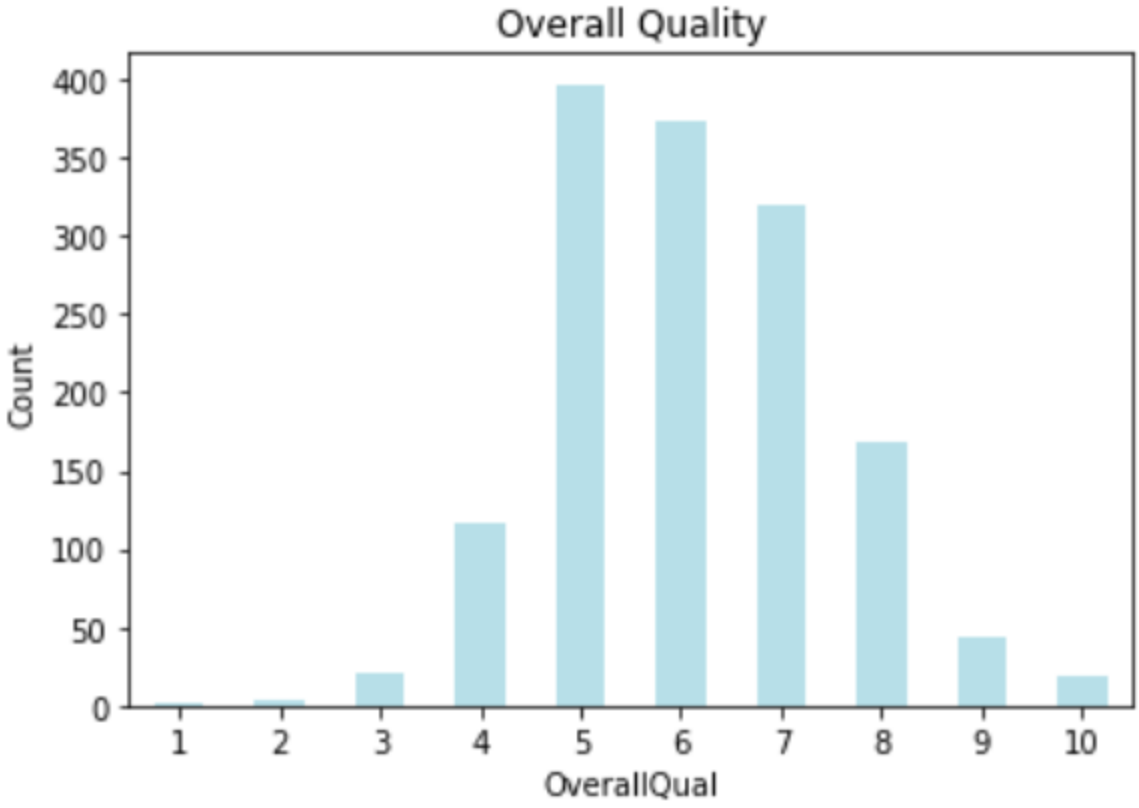


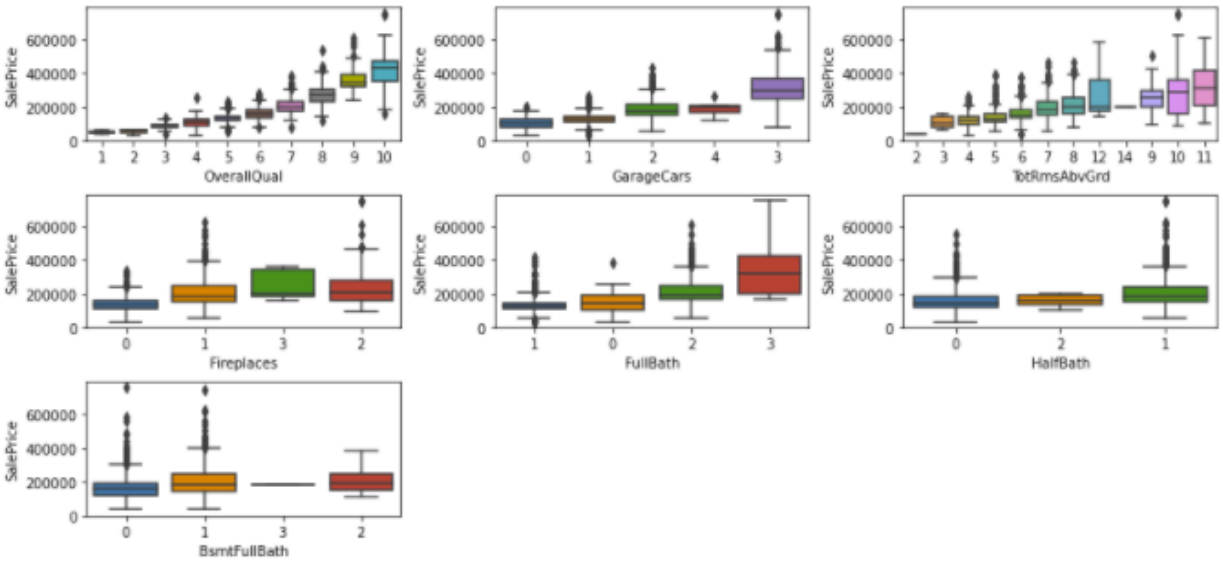
These two variables seem to be positively correlated with one another. There seem to be a few outliers as well which may affect the analysis.

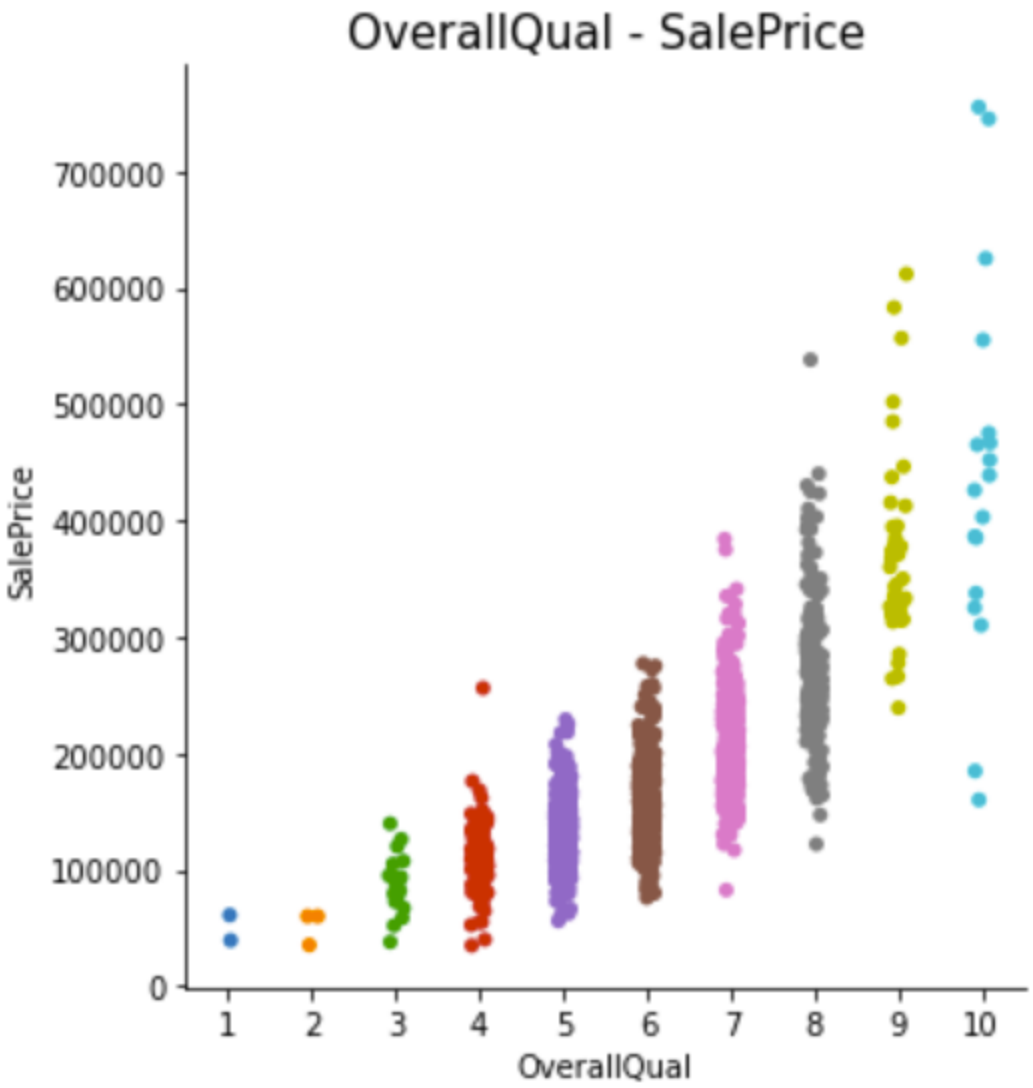
I wanted to create a heatmap so that I could visualize which features had the most correlation to the target variable:

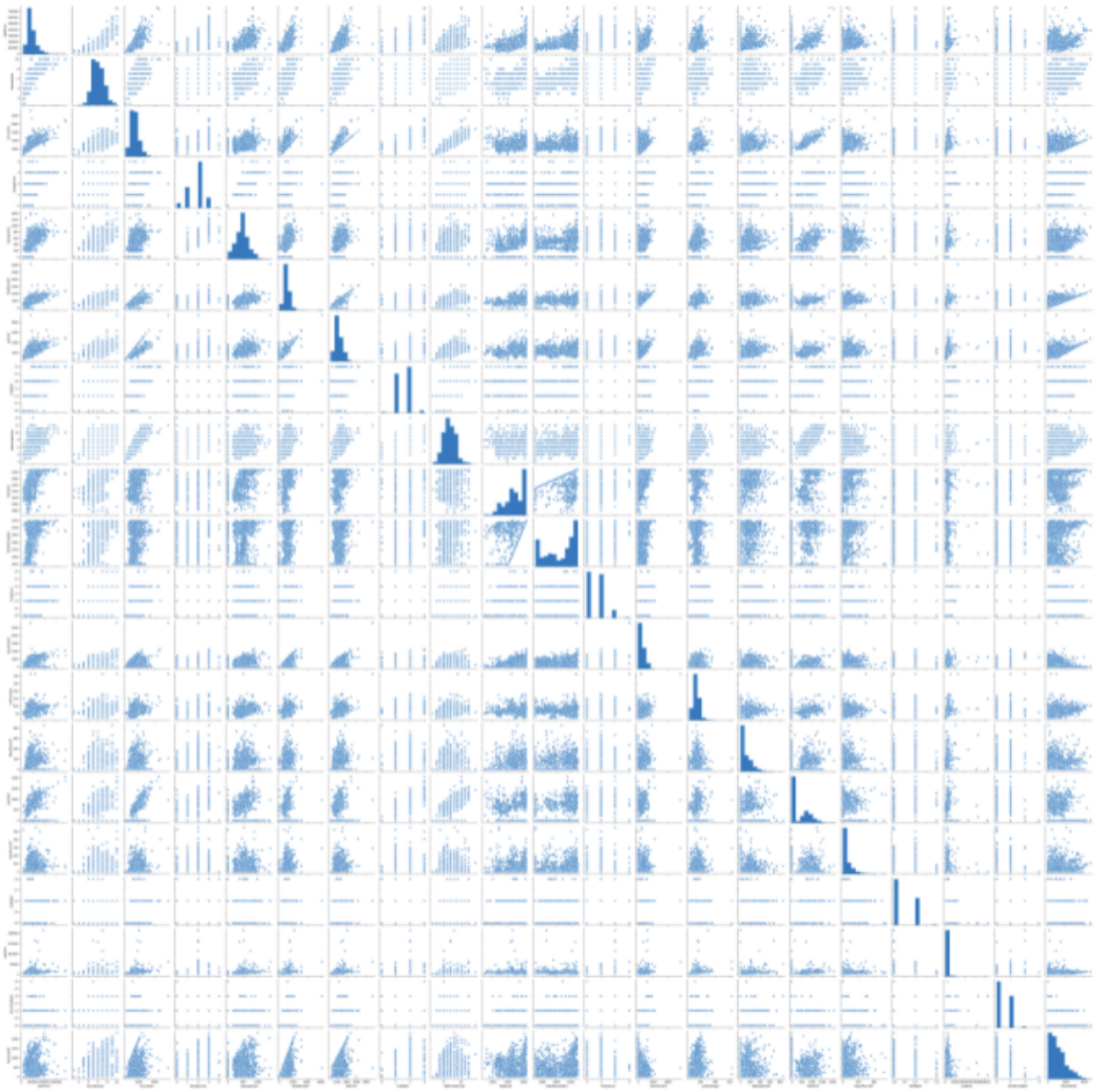


Based on the correlation heatmap, it appears that OverallQual (Overall Quality) has the most correlation to Sale Price. Also, the GrLivArea (Greater Living Area), GarageCars, GarageArea, TotalBsmtSF (Total square footage of the basement), 1stFlrSF (Total square footage of the first floor), and FullBath all have a relatively high correlation with Sale Price. I figured it would be a good idea to create barplots, catplots, boxplots, and pairplots of some of the relationships in order to visualize the relationship between the target variable and the other features.


Overall Quality
Count
OverallQual
0
50
100
150
200
250
300
350
400
1
2
3
4
5
6
7
8
9
10

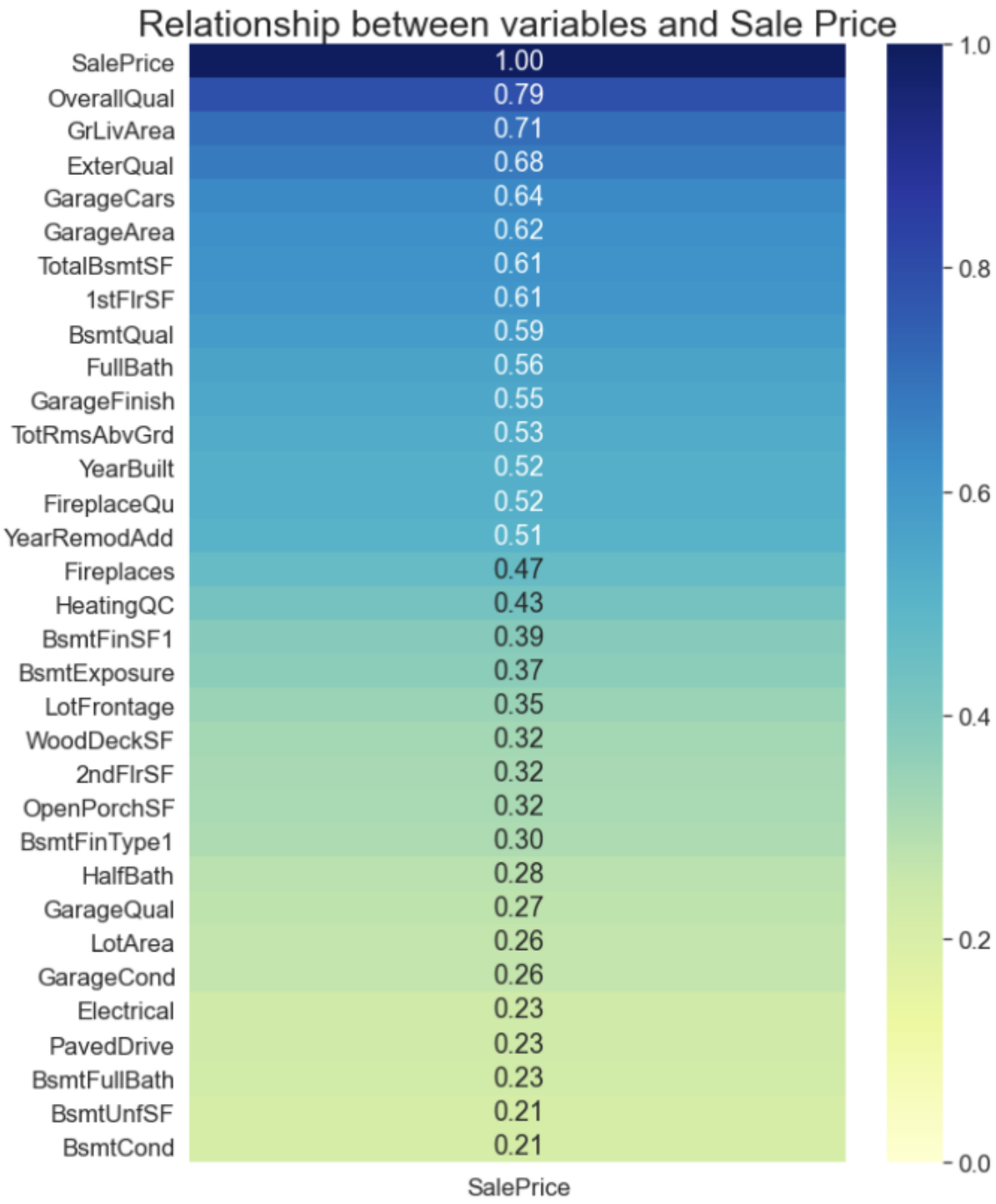

SalePrice
OverallQual
GarageCars
TotRmsAbvGrd
Fireplaces
FullBath
HalfBath
BsmtFullBath

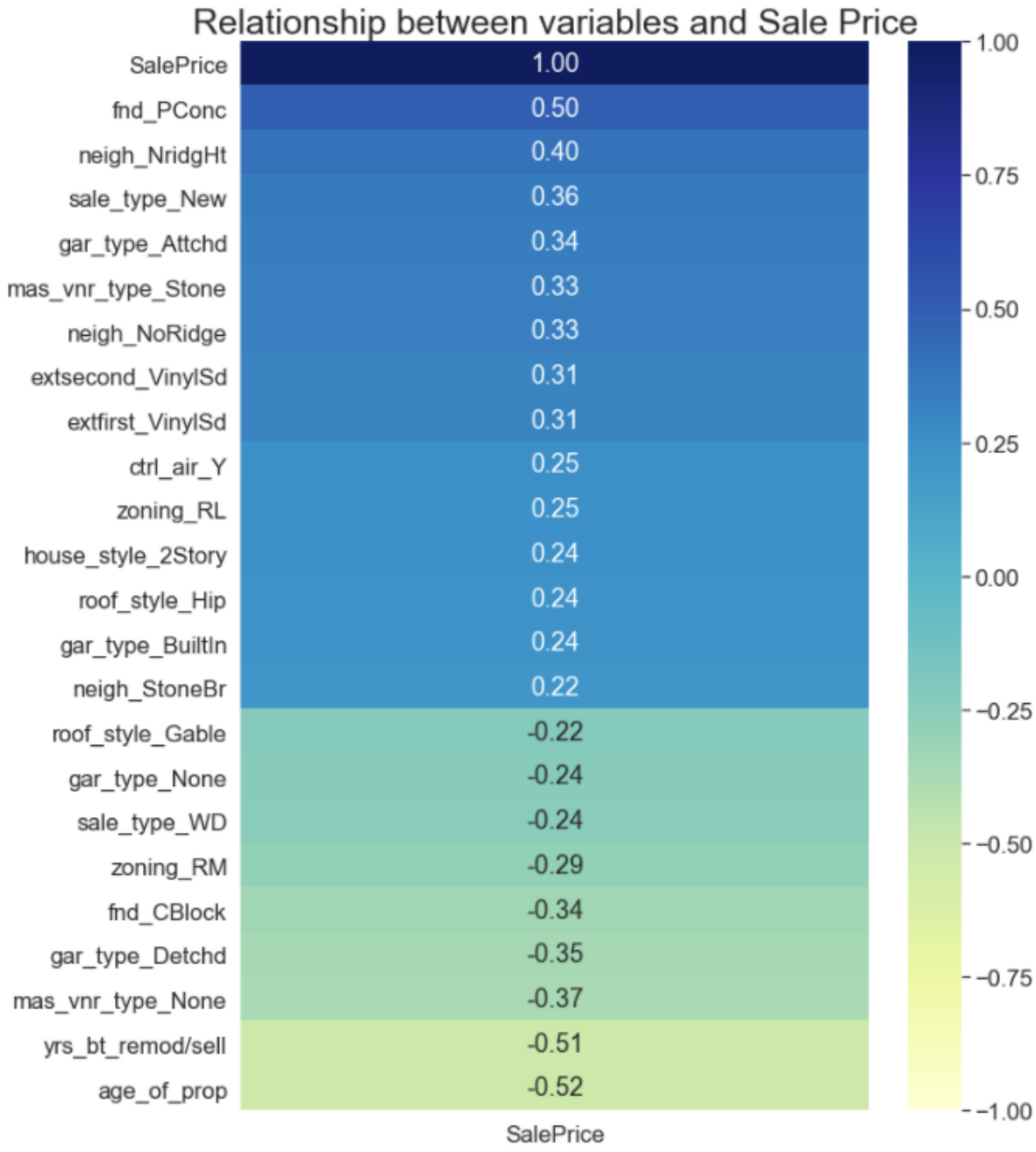

OverallQual - SalePrice
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
1
2
3
4
5
6
7
8
9
10
OverallQual



There are many features in this dataset so there are many relationships between variables. I found that, in performing an analysis, it is of extreme importance that I select the features that have the ability to influence the Sale Price.

I had another step to perform. I wanted to employ feature engineering by use of the 'get_dummies' feature and one-hot encoding so that the correlation table can take in consideration all of the features that may or may not have a positive or negative correlation to Sale Price. After I conducted the data engineering, I split the engineered features into two different groups. The heatmaps to the two groups are shown below:


Relationship between variables and Sale Price
SalePrice 1.00
OverallQual 0.79
GrLivArea 0.71
ExterQual 0.68
GarageCars 0.64
GarageArea 0.62
TotalBsmtSF 0.61
1stFlrSF 0.61
BsmtQual 0.59
FullBath 0.56
GarageFinish 0.55
TotRmsAbvGrd 0.53
YearBuilt 0.52
FireplaceQu 0.52
YearRemodAdd 0.51
Fireplaces 0.47
HeatingQC 0.43
BsmtFinSF1 0.39
BsmtExposure 0.37
LotFrontage 0.35
WoodDeckSF 0.32
2ndFlrSF 0.32
OpenPorchSF 0.32
BsmtFinType1 0.30
HalfBath 0.28
GarageQual 0.27
LotArea 0.26
GarageCond 0.26
Electrical 0.23
PavedDrive 0.23
BsmtFullBath 0.23
BsmtUnfSF 0.21
BsmtCond 0.21
SalePrice
1.0
0.8
0.6
0.4
0.2
0.0

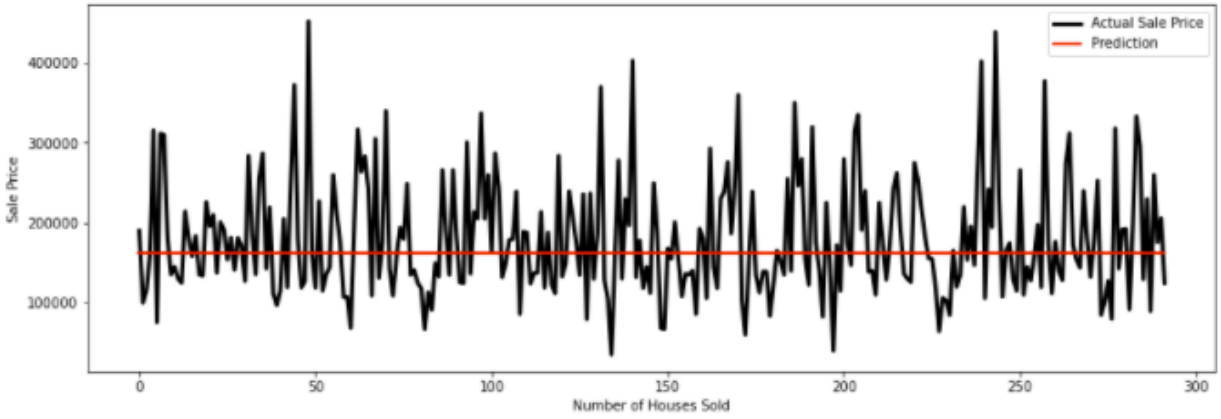


The updated heatmap shows features that have a higher correlation with the Sale Price than the previous analysis showed. Certain features are also negatively correlated. For instance, the age_of_prop feature (age of property) is negatively correlated to Sale Price which follows the logic that the older a property is tends to decrease the sale price of a house. Now that all of the features have been engineered and the features correlated to Sale Price have been analyzed, it was time to determine which model to use for the analysis.

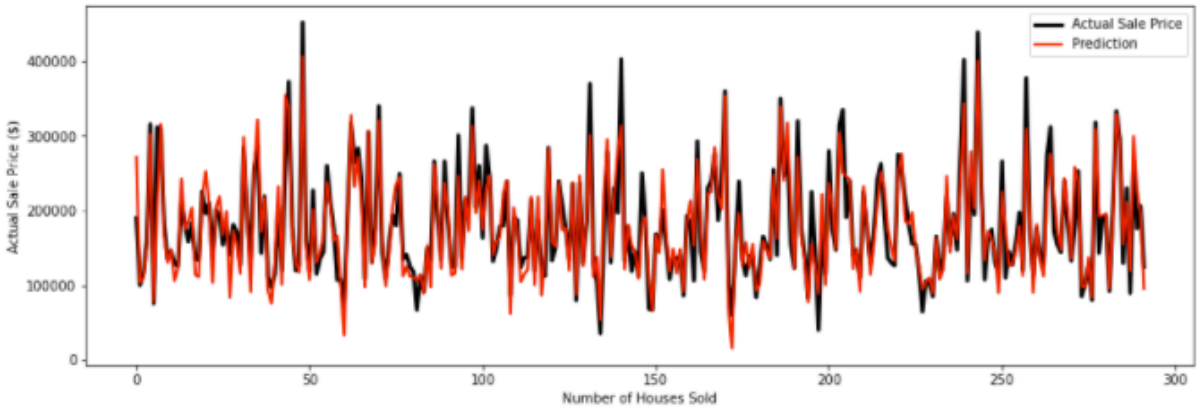
Model Selection

_____In my quest to find a model that will be effective, I decided to evaluate several models to the dataset and determine which regression model, after being trained, will perform best at predicting the sale price of homes.

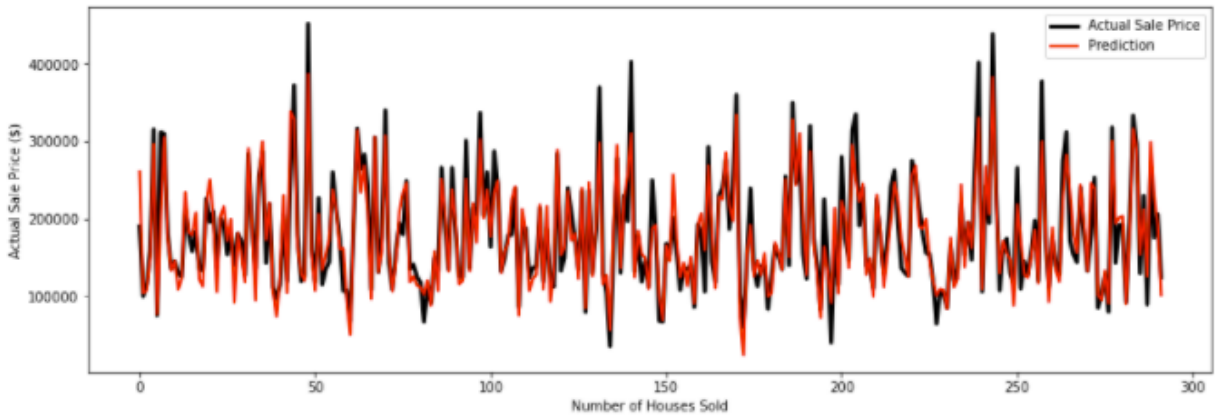
First, I used a dummy regressor in order to create a baseline for which to compare the results of my models to. This is useful because it helps me determine if any fine-tuning needs to be done to the model. After I trained each model on the dataset, I was going to compare the R2 values of the models against each other in order to determine which model performed the best. Below is a visualization of the dummy regressor against the dataset:



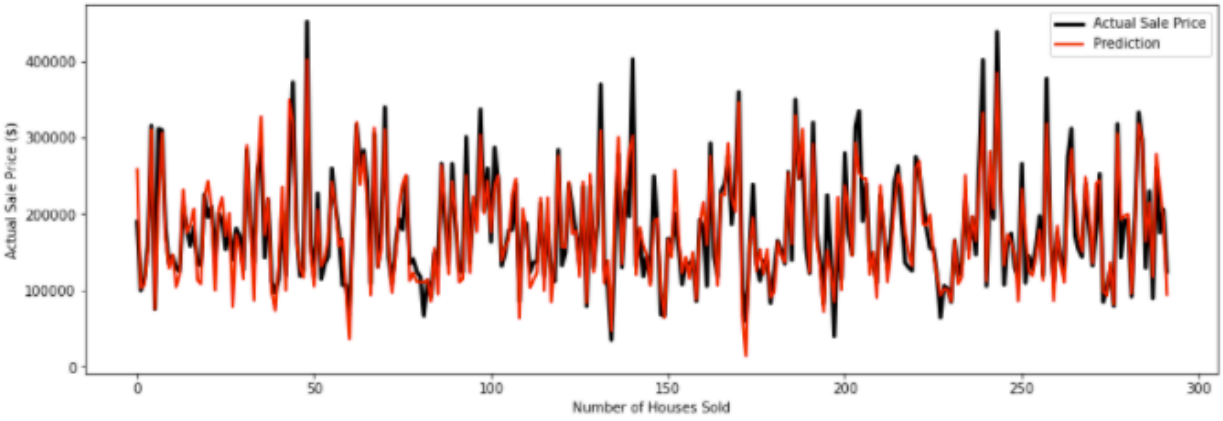
Next, I applied a Linear Regression model to the dataset:



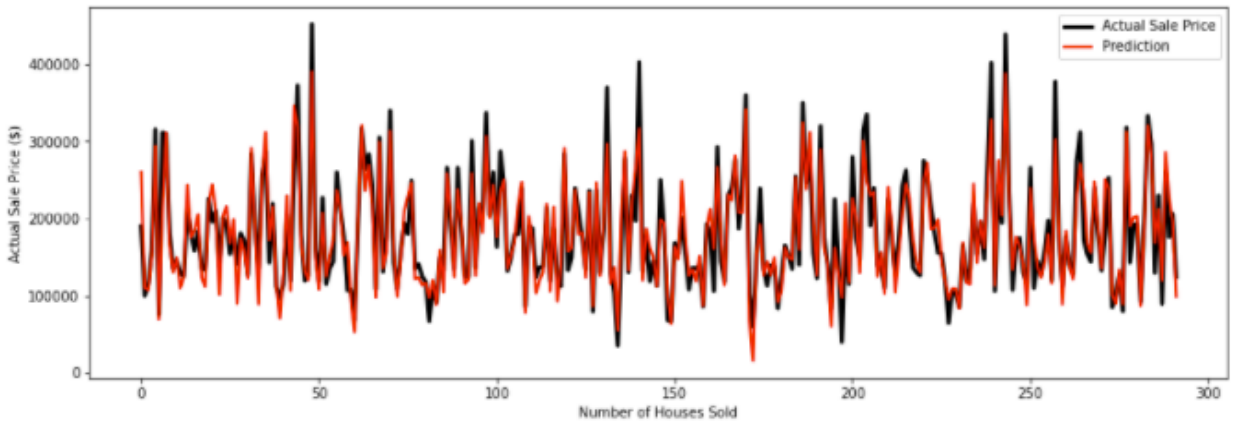
Next, I applied a Ridge Regression Model to the dataset. I tried to determine the best value of alpha using grid search cross validation. The visualization of the results are below:



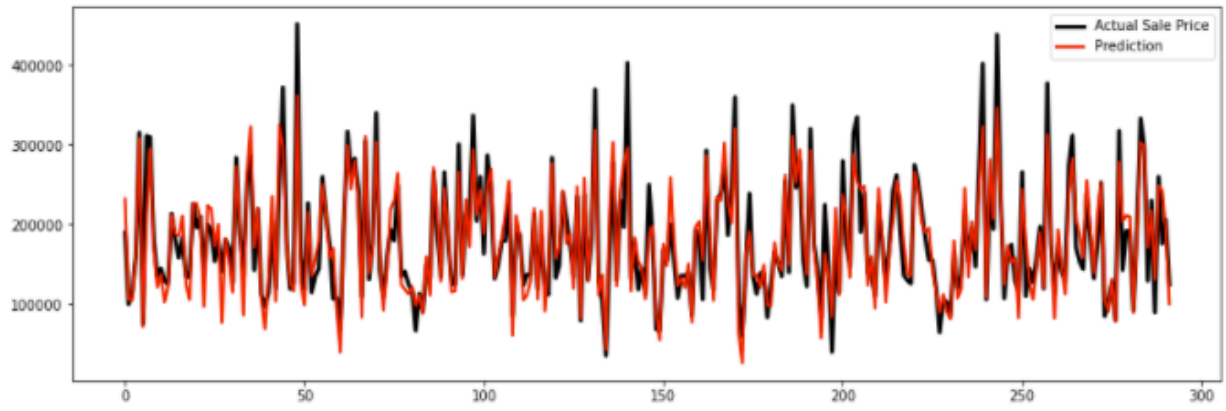
Next, I applied the Bayesian Ridge Regression model the the dataset. The resulting visualization is below:



Next, I applied the Lasso Regression model to the dataset. I conducted a grid search to determine the best value of alpha and according to the analysis, that was determined to be 1.



Finally, I utilized the Elastic Net Regression model. I also performed hyperparameter tuning and determined that the best value of alpha is .01.



The training data of all of the different models were cross validated for higher precision. The results of the 5 models are below and compared against the dummy regressor:

	Model	MSE	MAE	RMSE	10 Fold Cross Validation	Test R2
0	Dummy Regression	5,706,299,603.3013	56921.760274	75540.052974	-0.055243	-0.052565
1	Linear Regression	787,230,550.4122	20580.985219	28057.629095	0.795652	0.854790
2	Ridge Regression	738,256,890.5215	19787.737663	27170.883138	0.804858	0.863800
3	Bayesian Regression	781,625,158.6176	20762.556361	27957.559955	0.792811	0.855800
4	Lasso Regression	728,213,494.7433	19768.169275	26985.431157	0.800937	0.865700
5	Elastic Net Regression	813,270,906.9377	21507.399727	28517.905024	0.737231	0.849987

The following table shows the features that have the highest influence on sale price for each model:

	Linear Regression	Ridge Regression	Bayes Regression	Lasso Regreession	Elastic Net Regression
0	neigh_StoneBr	neigh_StoneBr	neigh_StoneBr	neigh_StoneBr	neigh_NridgHt
1	neigh_NridgHt	neigh_NridgHt	neigh_NridgHt	neigh_NridgHt	OverallQual
2	neigh_NoRidge	neigh_NoRidge	neigh_NoRidge	neigh_NoRidge	FireplaceQu
3	gar_type_None	OverallQual	OverallQual	OverallQual	ExterQual
4	extsecond_VinylSd	ExterQual	ExterQual	ExterQual	BsmtExposure

Takeaways

_____Based on the analysis, it appears that the models did fairly well in predicting the sale price of homes. Although its not an exact science as predictions never are, machine learning algorithms are useful and effective as they remove much of the human sensitivities and biases that have historically been present in housing market values.

The R2 value is a good indicator of the effectiveness of each model. The R2 value for the dummy regression model is -0.052565. However, for the 5 models, the R2 values range from 0.849987 - 0.863800. In comparison to the R2 value from the dummy regressor, all 5 models did great. Of course there is a margin of error for each model, but it seems as though the margin of error between the predicted values and the actual values of the Sale Price are not that high.

Each model, with the exception of the Bayes Regression Model and the Lasso Regression model, had a different output on which features were the most important and which features had the highest influence/correlation with Sale Price. However, it seems to be a consensus among all of the models that neighborhood plays a big factor in influencing Sale Price. This conclusion makes absolute sense because it is common sense that more affluent neighborhoods generally have higher sale prices for the homes than do houses in less affluent neighborhoods.

_____After performing this analysis, I think that regression models are a great tool in predicting the sale price of homes. This tool can be useful for home sellers, home

buyers, and banks who hire independent house appraisers. It could lead to more consistency within the housing market and it can remove the subjectivity of human knowledge and bias. As technology improves considerably, the tool may one day be able to replicate what humans are able to process and provide a more accurate assessment than a human is able to.

TAKEAWAYS

This project was highly informative and educational. I spent many hours trying to figure out the best approach to this assignment. I made mistakes and had to refine my code and thought process in order to achieve a much better performing model. I am satisfied with the performance of the models I selected. I think that there could have been some improvement and renditions done in order to increase my R2 score. Also, I learned the basics of model stacking and would have liked to include those algorithms in my assignment. Model stacking is an ensemble machine learning algorithm. "It uses a meta-learning algorithm to learn how to best combine the predictions from two or more base machine learning algorithms. The benefit of stacking is that it can harness the capabilities of a range of well-performing models on a classification or regression task and make predictions that have better performance than any single model in the ensemble."
https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/

I think that would have been a great concept for me to include in this project. However, we have not covered this topic in the course. I just think that it would have been effective as all of the models I implemented performed well. If given more time, I would have liked to attempt to integrate that into my code. In essence, I think that including model stacking in this project would have increased the overall performance of my models.

CONCLUSION

_____Based on the above analysis, the Lasso Regression model seems to have performed the best out of all comparable models. This conclusion is based on the fact that the R2 score for this model is 0.8657. Since this model has the highest metric for performance when compared to the other models, it is considered to be the most accurate model, relatively.