Final Report: Predicting Housing Prices

Patrick Oline

Problems Statement

Create a machine learning model that will help predict housing prices based on the most relevant features of a house.

- Focus is on housing data from Ames, Iowa.
- Goal is to inspect, clean, train and test an ML model.
- Success: 10-20% difference from the actual prices.

Data Wrangling

- Data from the kaggle.com website.

- The data has 79 explanatory variables.

- Drop any numeric/text columns with more than 5% missing values.

- Filled remaining missing information with the most common value.

Data Wrangling Continued

Created two columns.

1. ‘Built_to_Sale’ = difference between the sale year and the construction year

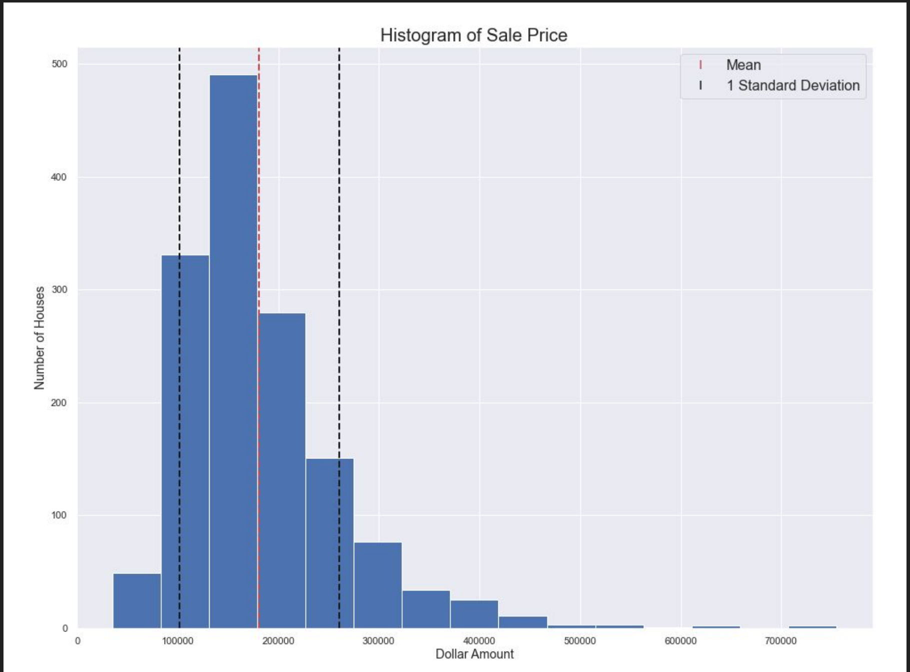
2. ‘Years_Since_Remod’ = number of years it has been since remodeled

Dropped columns because they were no longer needed, irrelevant, or leaked data:

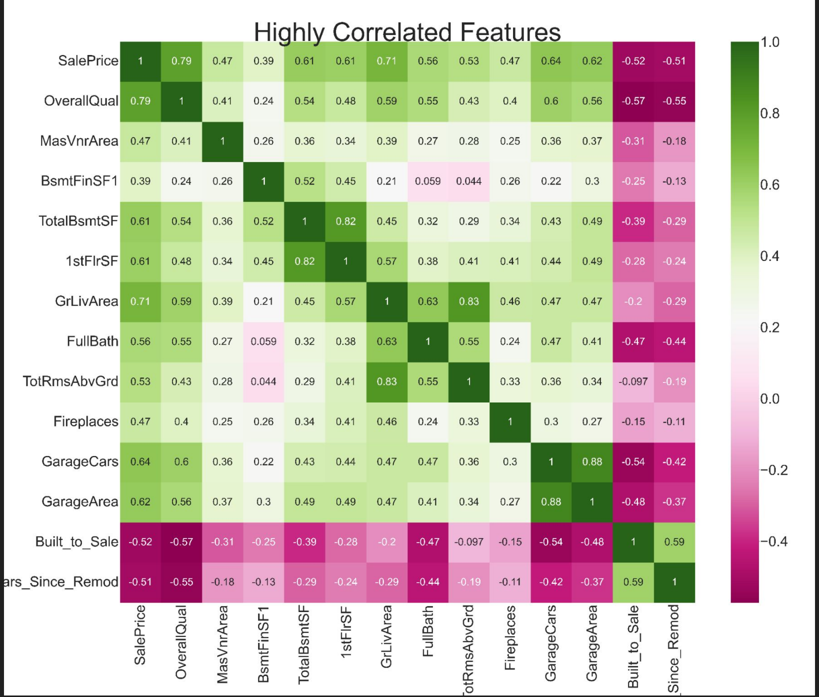
[‘Id’, ‘YearBuilt’, ‘YearRemodAdd’, ‘MoSold’, ‘SaleCondition’, ‘SaleType’, ‘YrSold’]

Exploratory Data Analysis

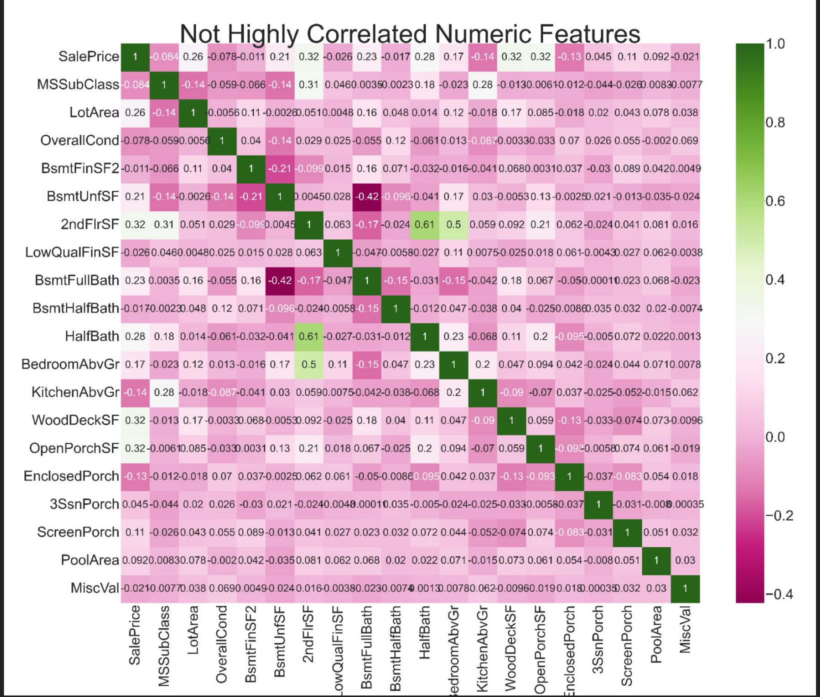
1. Mean = $180,921.20
2. Standard Deviation = $79,442.50
3. Mean±STD = ($101,478.70, $260,363.70)
4. Skewed to the right.



Numeric Feature Analysis

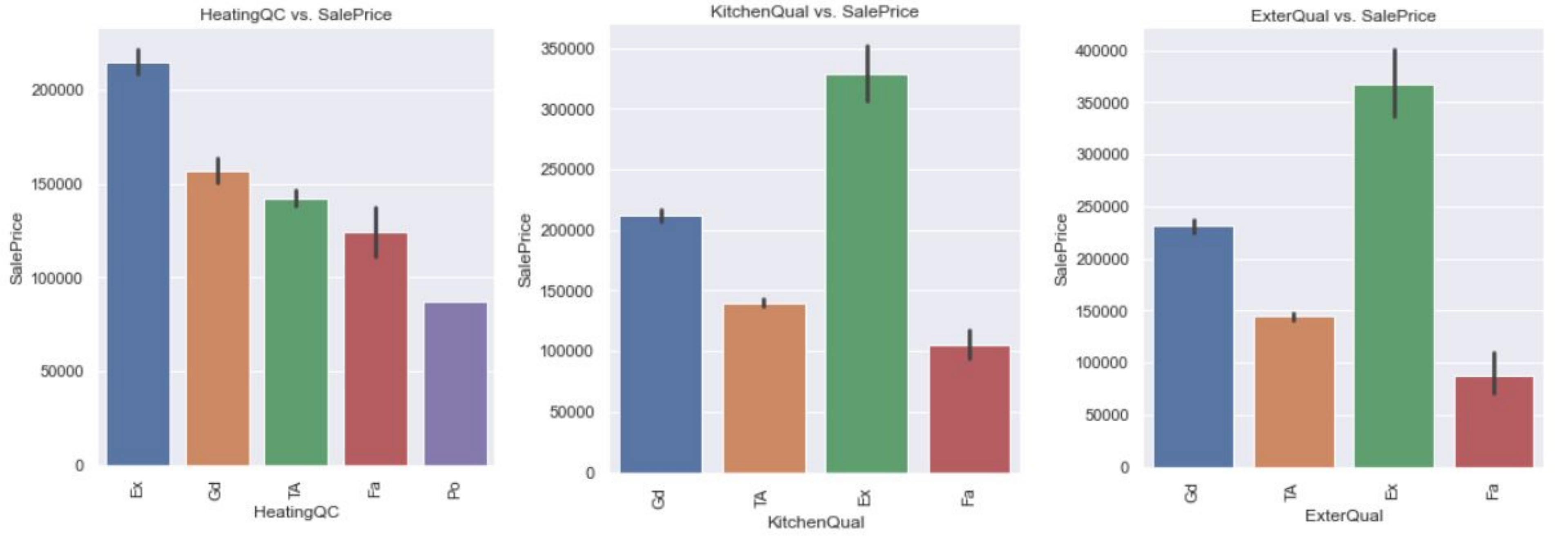


Correlation to ‘*SalePrice*’ > 0.39 or < -0.5



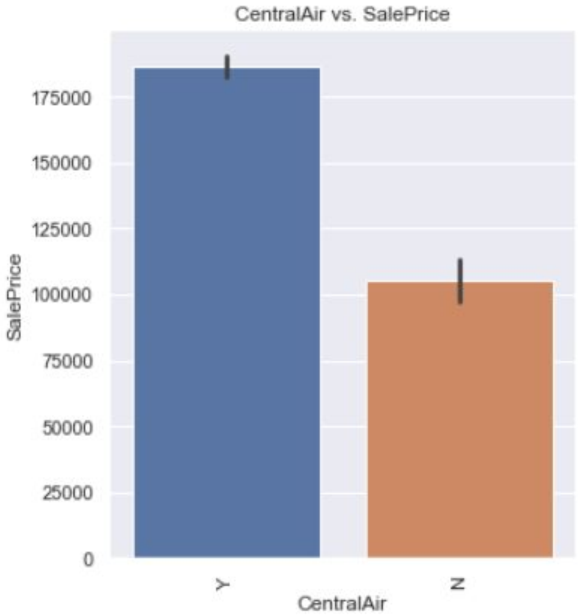
Remaining numeric columns

Categorical Feature Analysis

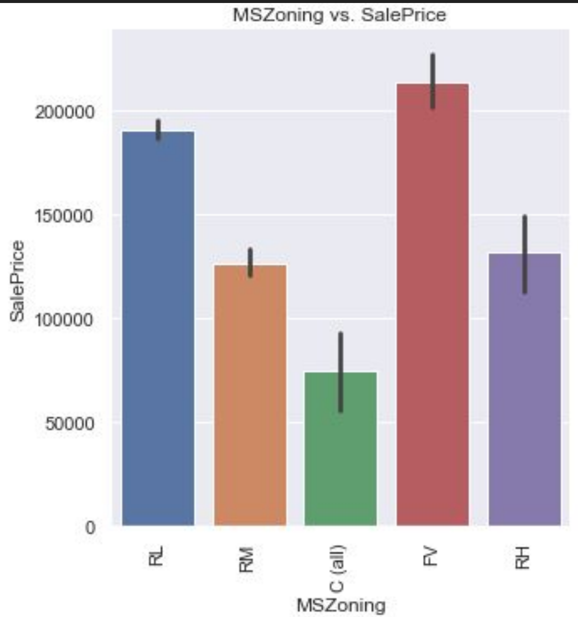


Ex = Excellent Gd = Good TA = Average/Typical Fa = Fair Po = Poor

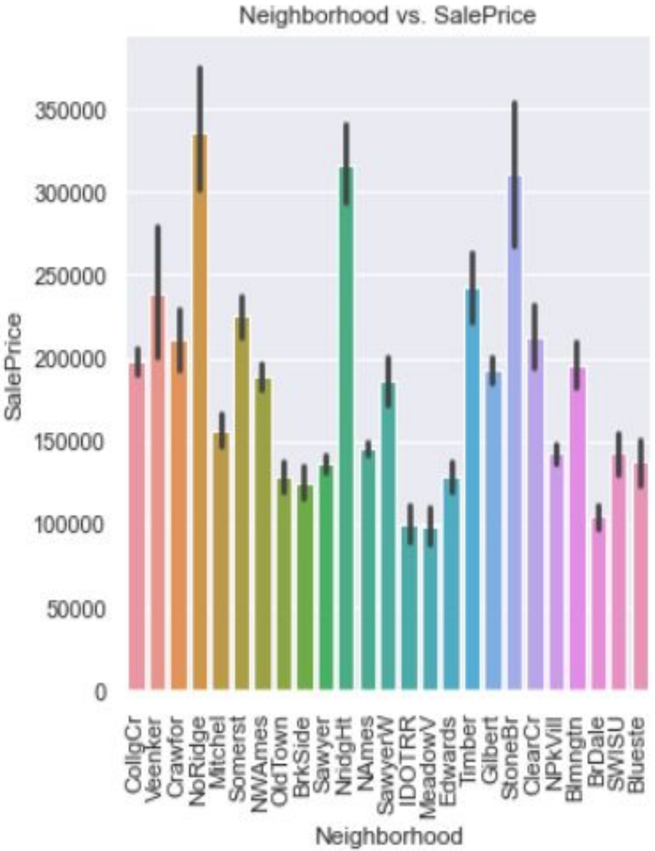
Categorical Feature Analysis



Y= Yes
N=No



RL = Residential Low
RM = Residential Medium Density
RH = Residential High Density
C(all) = Commercial
FV = Floating Village Residential



Model Selection

Linear Regression Model

1. Standard LR Model
2. Cross Validation
3. Correlation Threshold

Random Forest Model

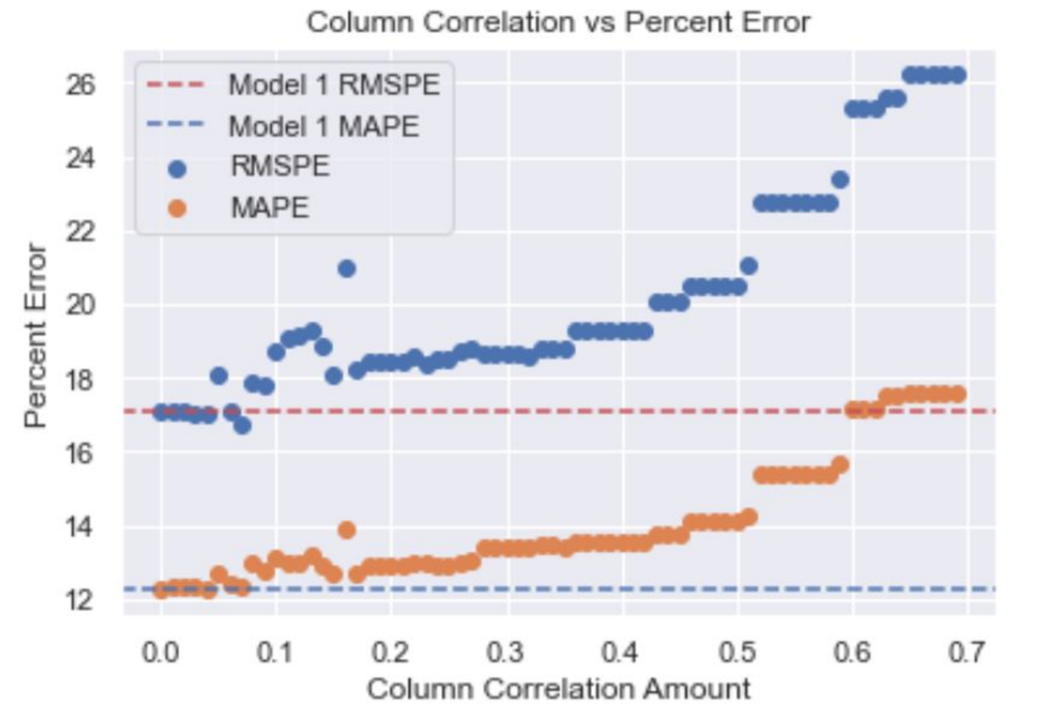
1. Standard RF Model
2. RF Random Search CV
3. RF Grid Search CV

Linear Regression Models

1. Standard LR Model
    a. RMSPE = 17.09%
    b. MAPE = 12.30%

2. Cross Validation
    a. Unknown Error

3. Correlation Threshold
    a. RMSPE = 16.73%
    b. MAPE = 12.25%



Random Forest Models

Standard RF Model

RMSPE = 18.73% MAPE = 10.67%

RandomSearchCV	GridSearchCV
n_estimators: 400 min_samples_split: 2 min_samples_leaf: 1 max_features: ‘sqrt' max_depth: None bootstrap: False	n_estimators: 200 min_samples_split: 8 min_samples_leaf: 3 max_features: 3 max_depth: 90 bootstrap: True
RMSPE = 18.93%	**RMSPE =** 27.32%
MAPE = 10.50%	**MAPE =** 15.58%

Comparing Results

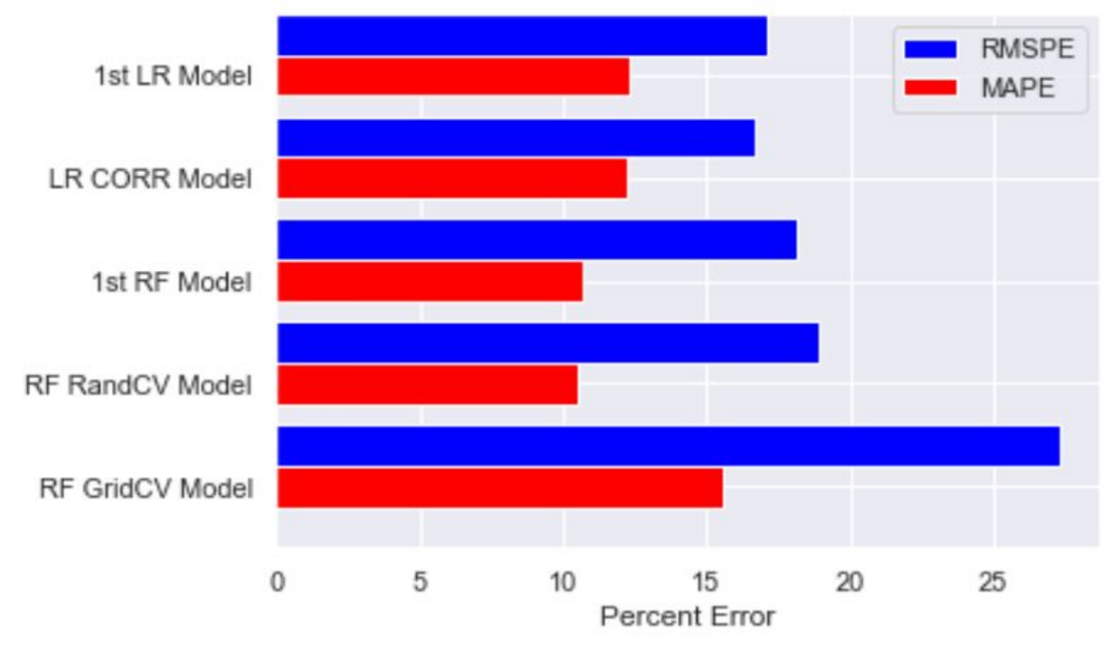
Name	RMSE	RMSPE	MAE	MAPE
1st LR Model	$34044.04	17.09%	$21260.01	12.30%
LR CORR Model	$34015.87	16.73%	$21254.74	12.25%
1st RF Model	$27763.13	18.18%	$17505.46	10.67%
RF RandCV Model	$28930.04	18.93%	$16803.02	10.50%
RF GridCV Model	$42705.14	27.32%	$24293.05	15.58%

Takeaways

- Linear regression model, ideal to remove a few low correlated features.

- The mean absolute error in the RF models was lower than the LR models.

- RF model produced decent RMSE in 2 of the 3 models.



Predicting Sample Prices

Using the optimal random forest model, I collected random data from the training set for each of the columns and ran a prediction for three different house sale prices.

MSSubClass	**LotArea(sqft)**	**OverallQual**	**KitchenQual**	**Remodeled**	**Built to Sale**	**SalePrice**
20	14000	4	Good	3 years	47 years	159,153.19
50	9135	6	Average	1 years	31 years	161,670.08
20	27650	5	Average	0 years	5 years	167,280.12

MSSubClass
20 = 1-STORY 1946 & NEWER ALL STYLES
50 = 1-1/2 STORY FINISHED ALL AGES

Future Research

- Collect more data to see changes over time.
- Further investigating into local locations like neighborhoods.
- Investigate how the model applies in other locations around the country.
- Increase the accuracy of the model.
- Try other ML algorithms.
- Investigate collinearity by finding and removing the features with high collinearity.