Predicting House Prices in Ames, Iowa

Machine Learning Project

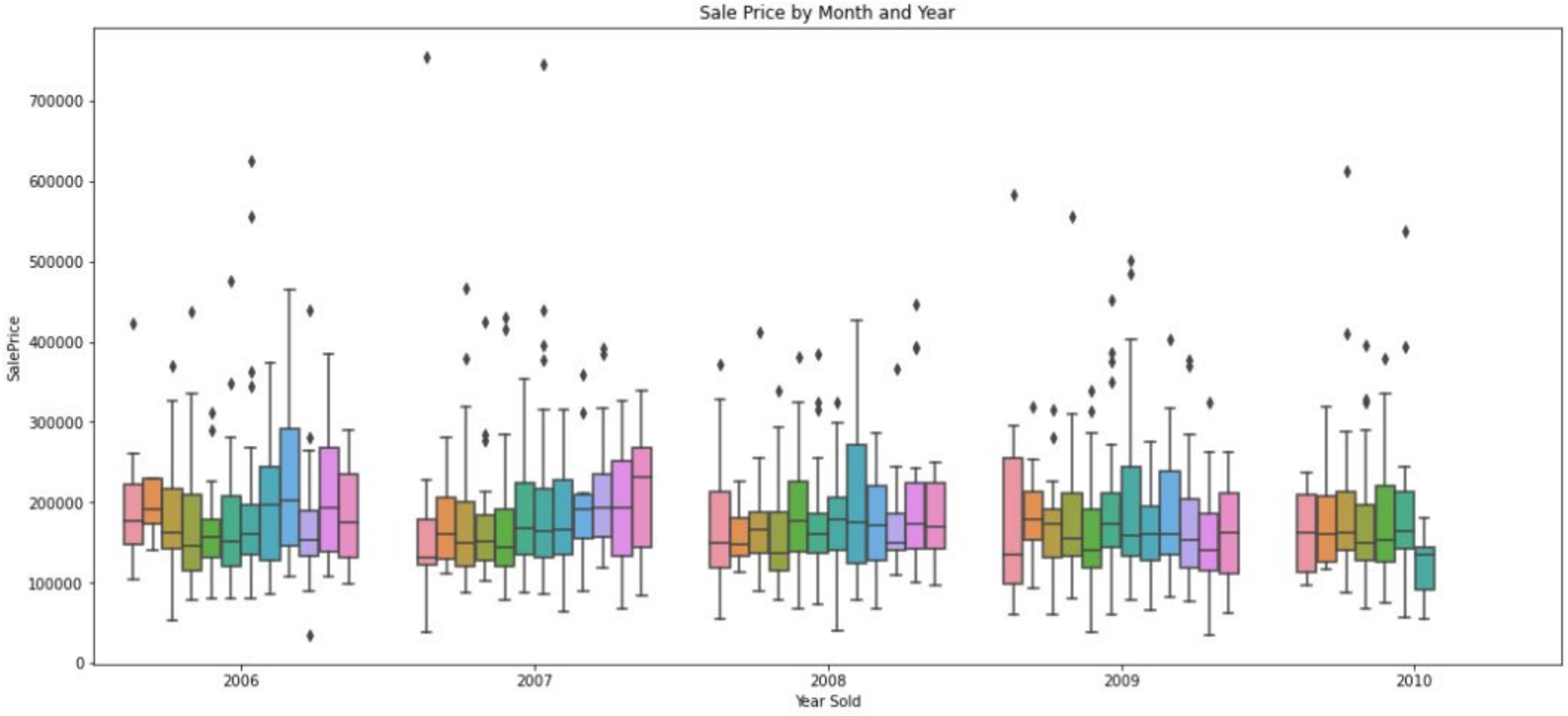
Background

- Kaggle dataset with information on 2,930 residential homes in Ames, Iowa (home of Neva Morris, the oldest living person in the US who died in 2010 at age 114). There are 79 explanatory variables.
- Objective is to fit a machine learning model that will best predict 1,459 house prices and identify the most important features for these predictions

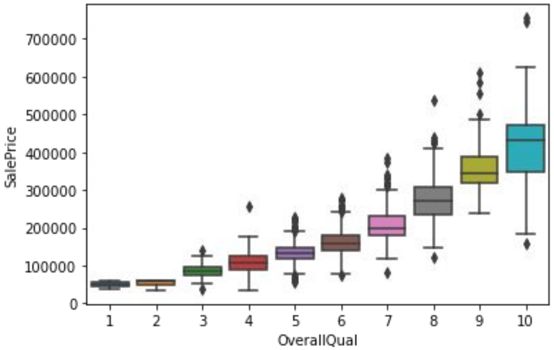
Outline

Exploratory Data Analysis	Data Cleaning & Pre-Processing	Model Fitting & Evaluation
• Data visualization to identify trends and prepare for data cleaning	• Imputing missing values • Feature selection/ de-selection • Feature engineering	• Comparing RMSE across models to determine the best fit for final predictions

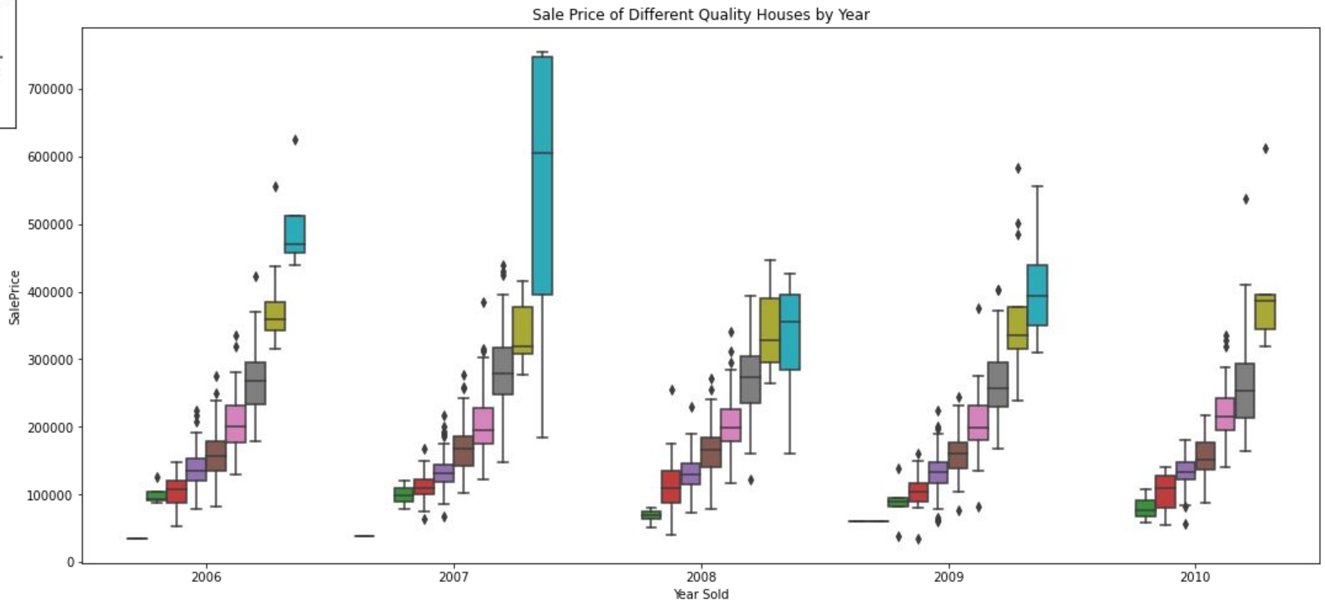
EDA- Pricing Seasonality



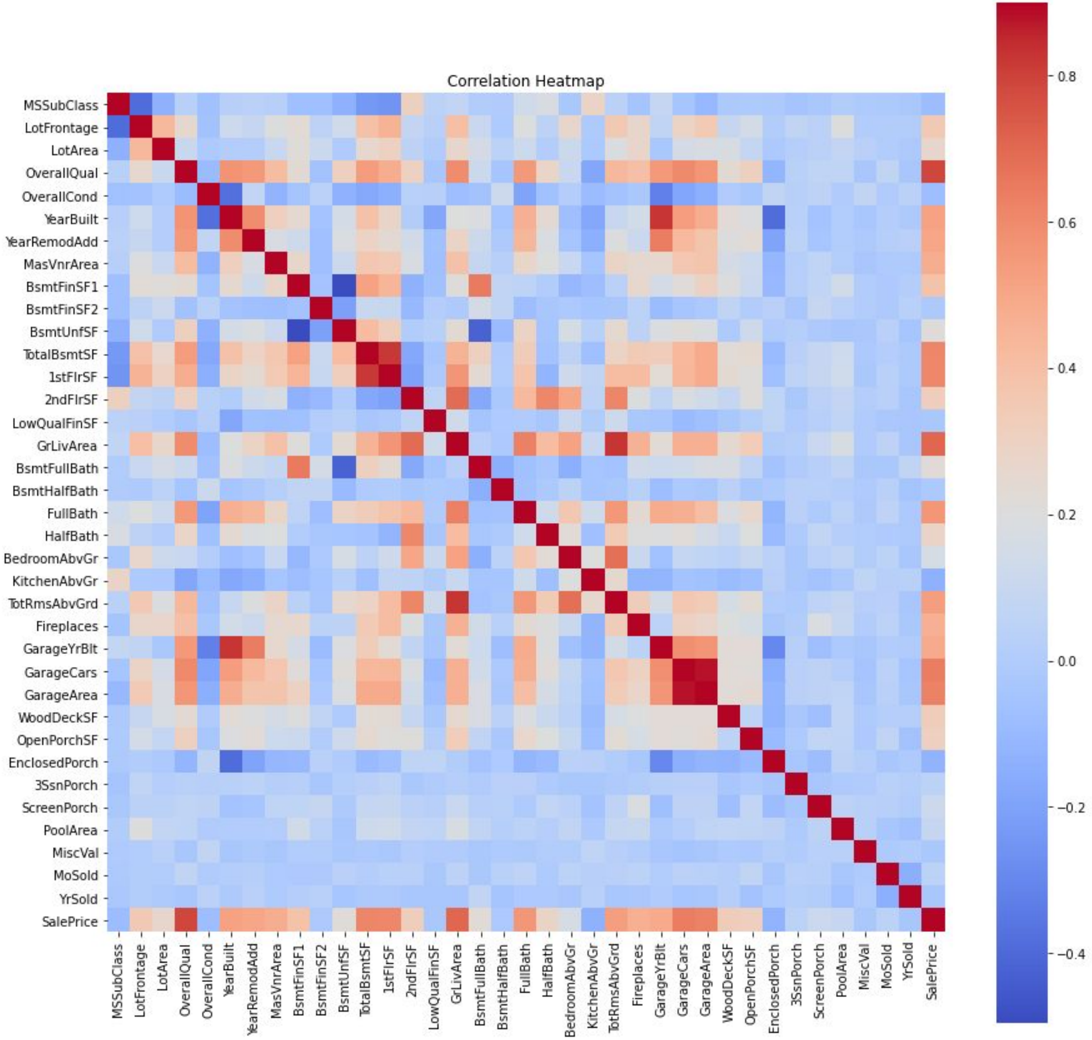
EDA- Sale Price & Overall Quality



Did higher quality houses decrease in price during the 2008 housing crisis?

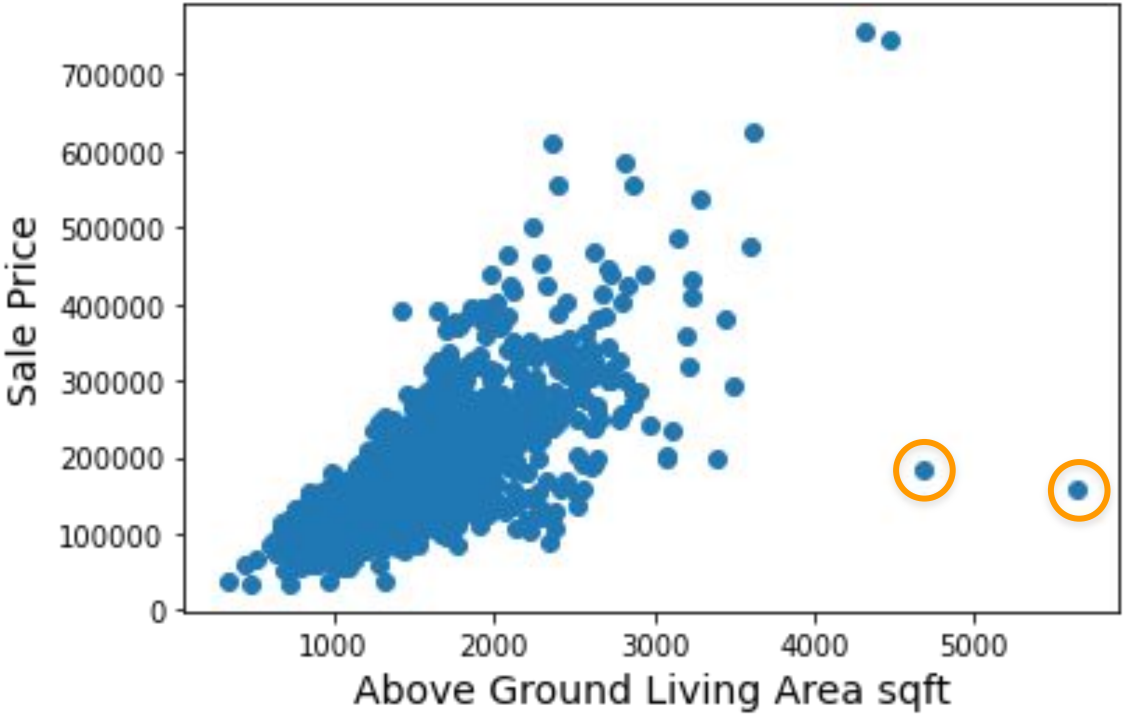


EDA



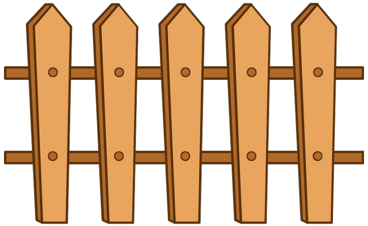
- Drop **GarageCars** since it's highly correlated with GarageArea
- Drop **GarageYrBuilt** since it correlates highly with YearBuilt
- Drop **TotRmsAbvGrd** (corr to GrLivArea)
- Drop **1stFlSF** + **2ndFlSf** because of GrLivArea and the close correlation between basement and first floor sq ft

Data Cleaning- Eliminating Outliers



Data Cleaning- Imputing Missing Values

20 out of 29 columns with missing values were related to non-essential, or "bonus" home features and easily imputed with zero or "none"



Data Cleaning- Imputing Missing Values

- For the remaining columns with missing values, 7 were categorical variables which were imputed with mode
- **LotFrontage** (16.7% missingness) was imputed with neighborhood mode
- **Utilities** was dropped, as all values were the same except for one in the train dataset

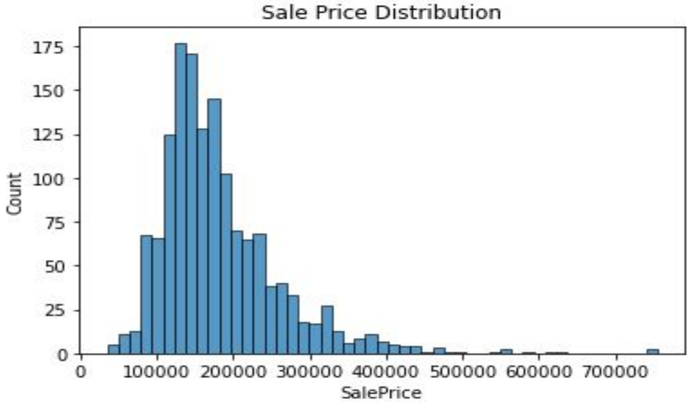
Feature Engineering- New Variables

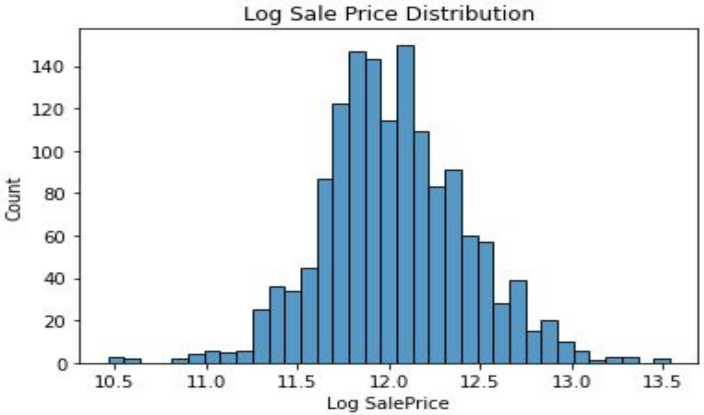
- **HouseAge** was calculated from **YearSold** and **YearBuilt** to replace **YearBuilt**
- **YearsSinceRemod** was calculated to replace **YearRemodAdd**
- **TotalBathrooms** was calculated to consolidate basement and non-basement full bath and half bath variables
- **PorchSF** was calculated to consolidate the square footage of various porch types
- **MSSubClass**, **MoSold** (month), and **YearSold** were changed from numeric to categorical

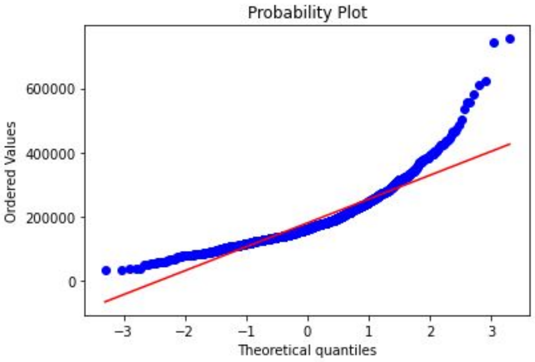
Encoding Categorical Variables

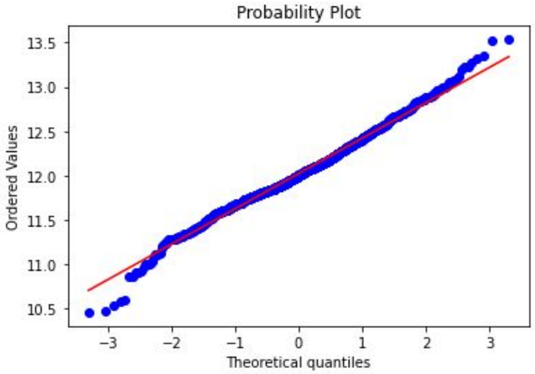
- Ordinal Encoding was used to encode 14 categorical variables, such as for various quality scores
- Non-ordinal categorical variables were dummified for linear models

EDA- Dependent Variable

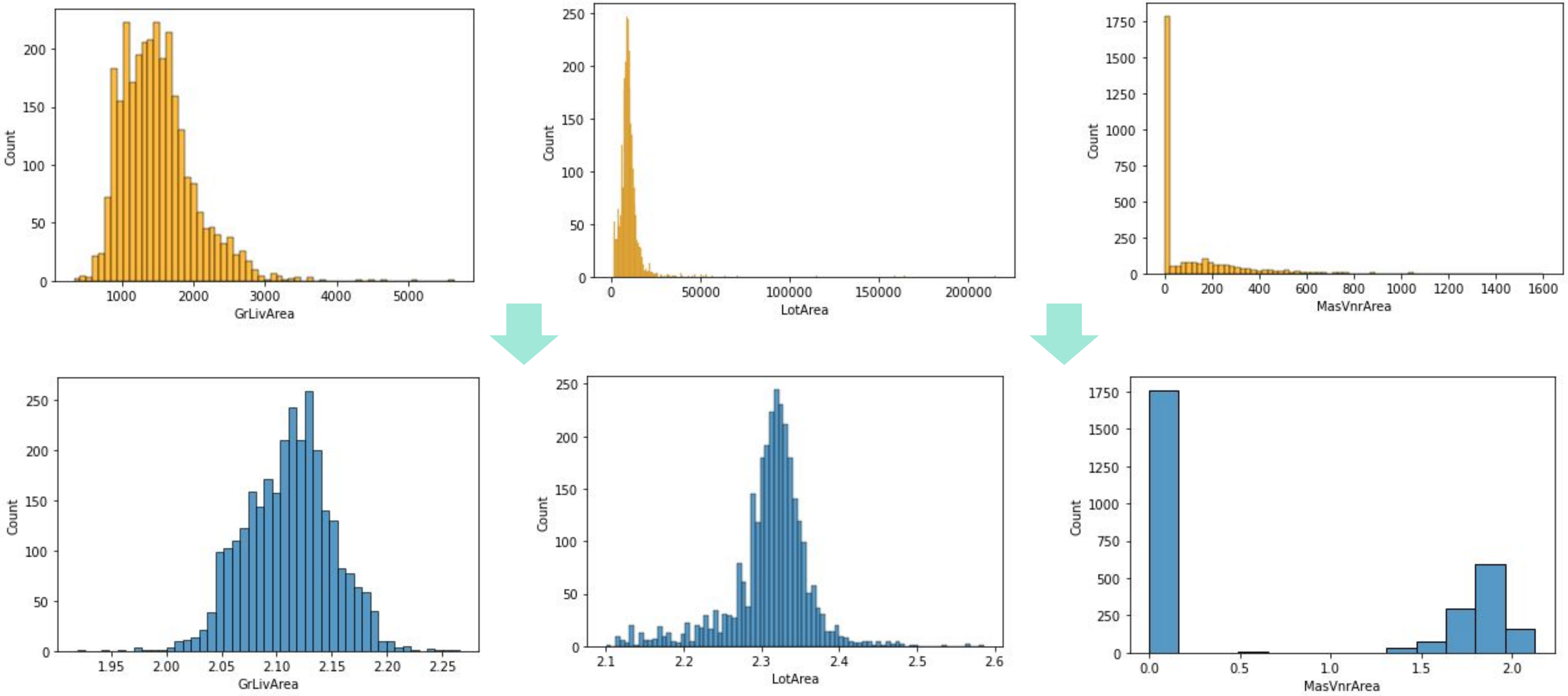








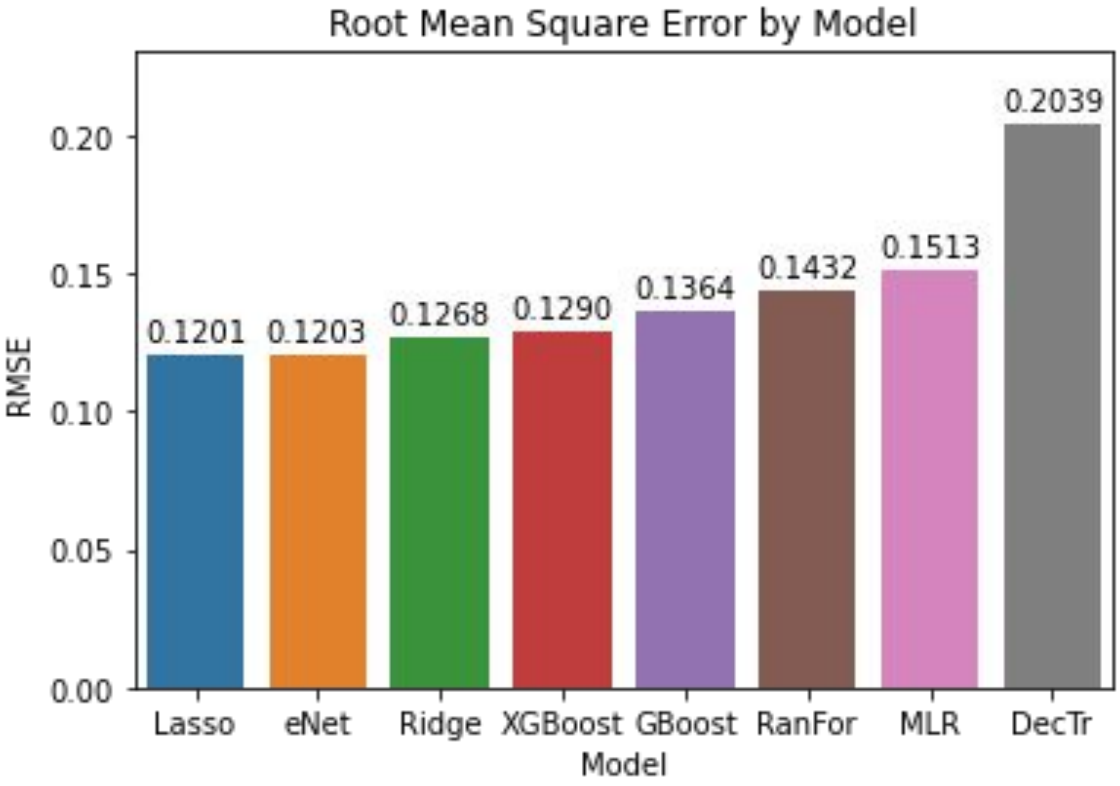
Log Transformation of Highly Skewed Variables



Model Fitting & Evaluation

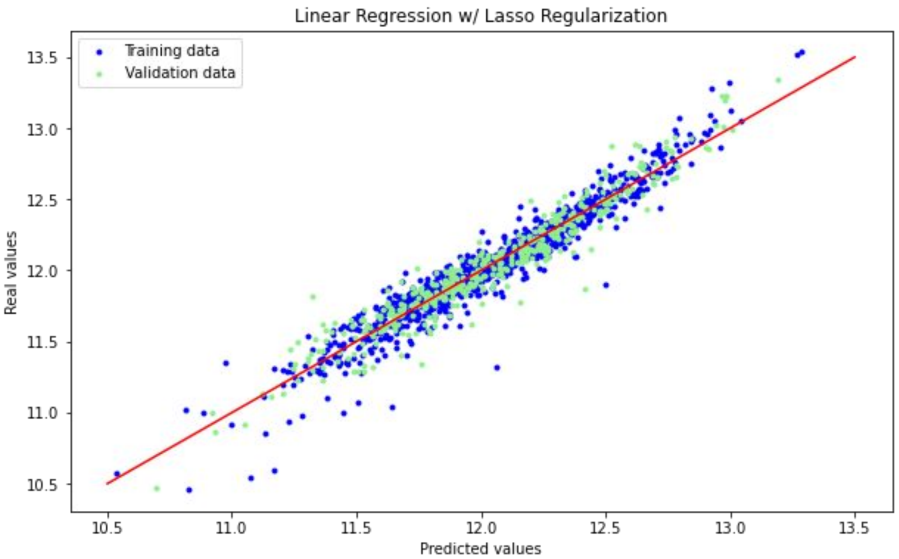
- Multiple linear regression
- Penalized linear regression
  - Ridge
  - Lasso
  - Elastic Net
- Decision Tree
- Random Forest
- Gradient Boost
- XGBoost

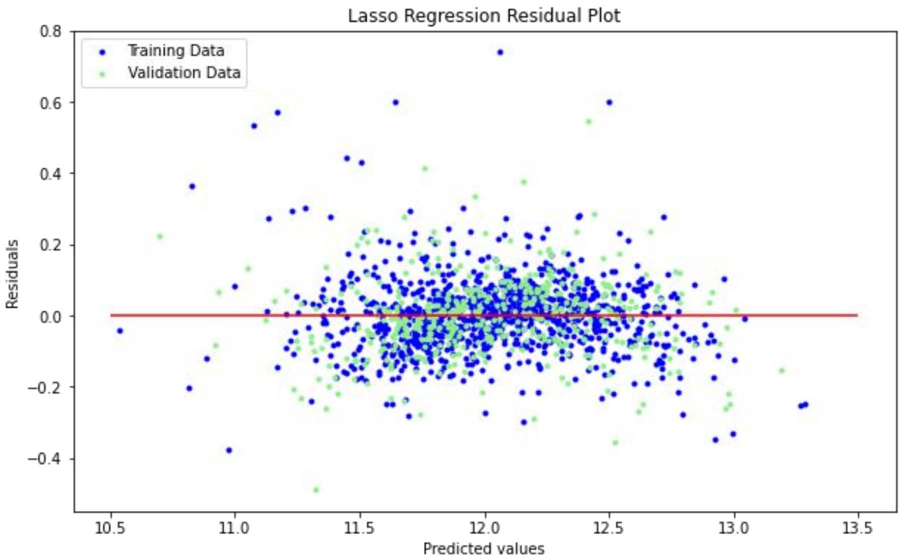
Model Comparison



- Data was split into train and test sets and evaluated using 10-fold cross validation
- Lasso penalized linear regression delivered the lowest RMSE
- XGBoost was the highest performing non-linear model; however its RMSE was still higher than the penalized linear models after using a grid search to tune parameters

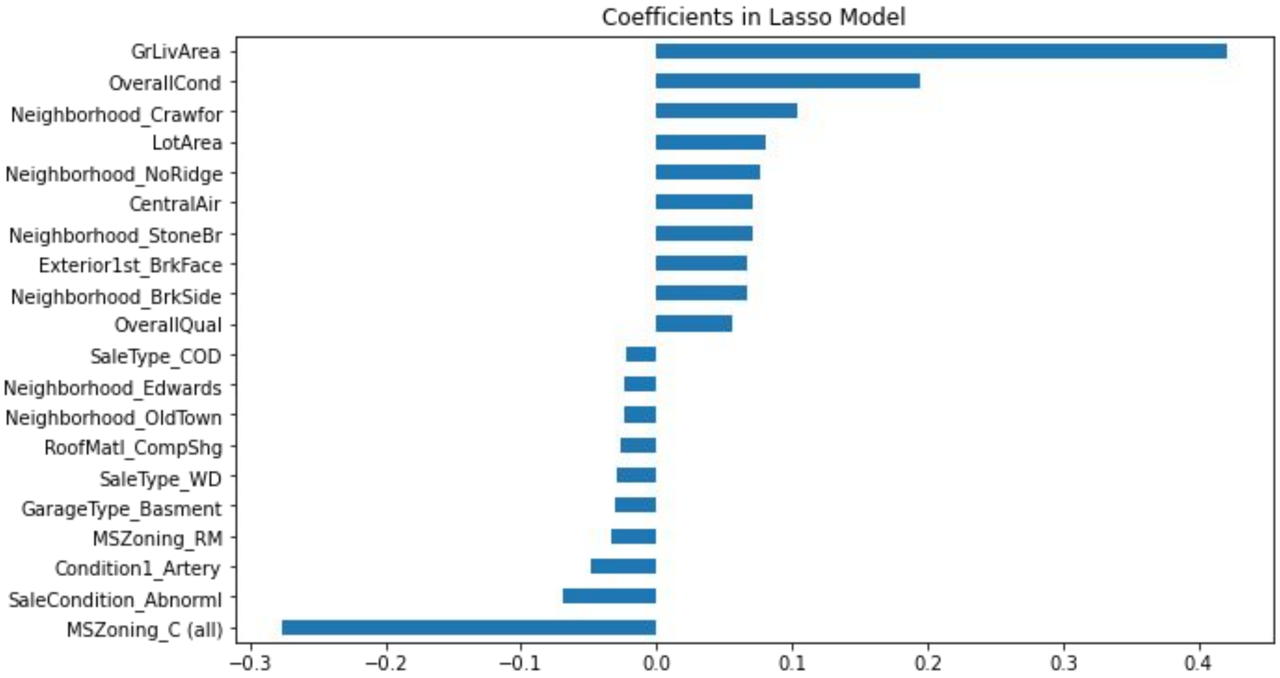
Lasso Model- Prediction Plot & Residuals





Visually, the regression seems to be a good fit and the residuals are randomly distributed around the center line with no clear pattern

Lasso Model- Top Coefficients



- **Above ground living area** is the most significant predictor of sale price, followed by **overall condition**
- Six of the top coefficients relate to **neighborhood**. Houses in Crawford tend to be the most expensive
- **Commercial zoning classification** was the top negative coefficient followed by an abnormal **sale condition** indicating a trade, foreclosure, or short sale

Key Takeaways

- A bigger house is not always a better house, but in Ames it is most likely a more expensive house
  - Home owners can significantly increase the value of their home through constructing home extensions, when possible
- Home buyers looking to save money on a nicer house should consider neighborhoods like Edwards and Old Town
- And finally, indecisiveness is a tough trait to have when it comes to machine learning

Opportunities for Further Analysis

- The final predictions scored in the top 25th percentile on Kaggle. Given additional time and resources, fine tuning some of the model parameters may provide better results as we find an ideal balance between overfitting and underfitting
- We may also try blending or stacking models, although this approach will add a layer of complication and make the model more difficult to interpret
- Testing out different adjustments to feature engineering, such as binning categorical variables (neighborhood, month sold) or performing different transformations of skewed variables may also improve the model predictions