Linear Regression Modeling of Ames Housing Data

General Assembly Project Two

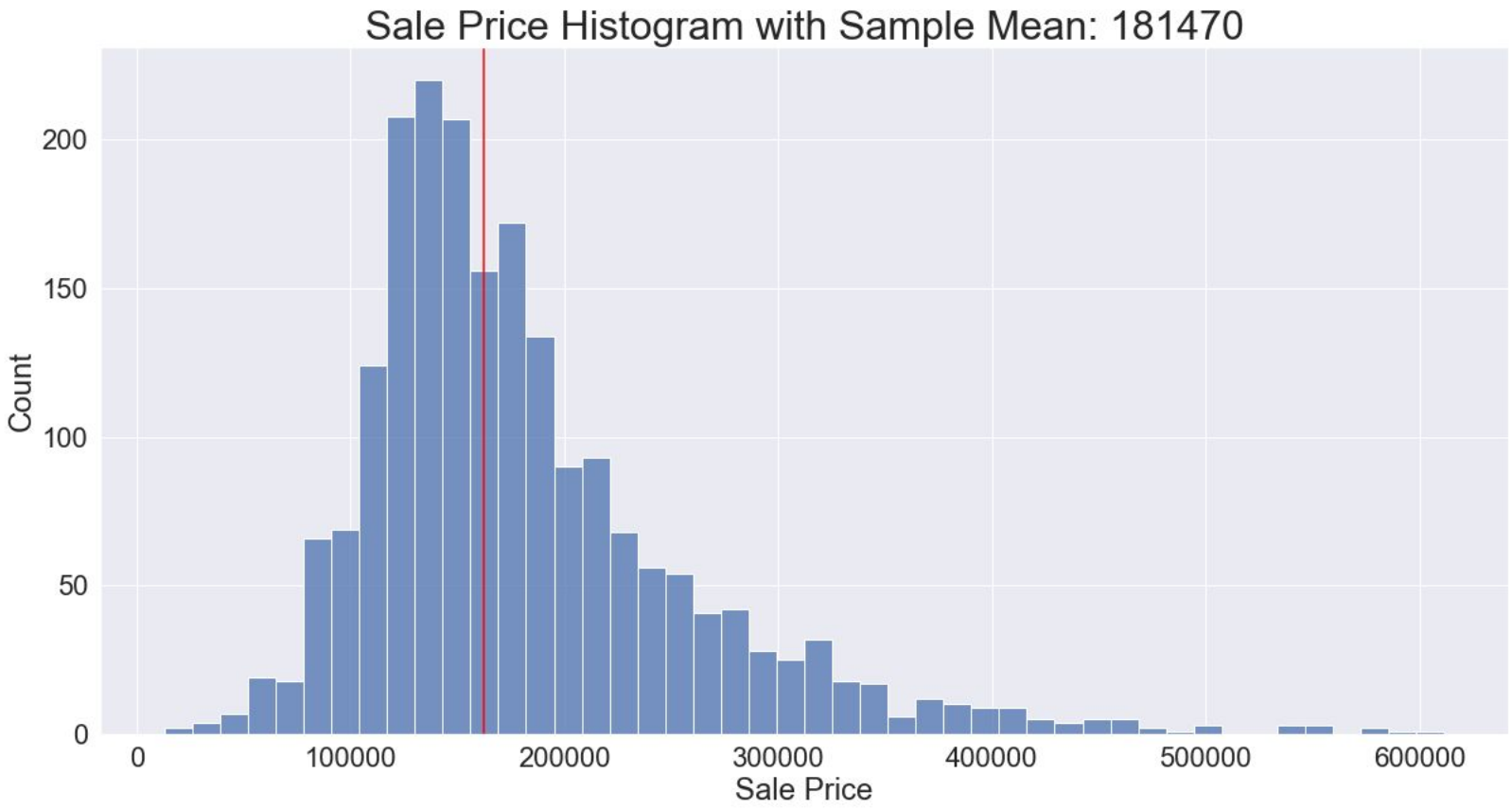
Peter Wentzel

Problem Statement and Data Set

- The goal of this study was to create a linear regression model from data provided by the assessor’s office in Ames, Iowa to predict home sale prices
  - How does the model compare against the null model (data set mean)?
  - Does the model provide further insight into home features that might add significant value to property for realtors or new home construction.
- A training set was used to predict sale prices of homes and then using the model on the testing set, the predicted values were compared against the known sale prices for a Kaggle competition
  - The data was collected was from 2006 to 2010 home sales
  - Training set entries of 2051 and test set entries of 878, 81 columns of features with the sale prices omitted from the test set

General Data Trends and Findings

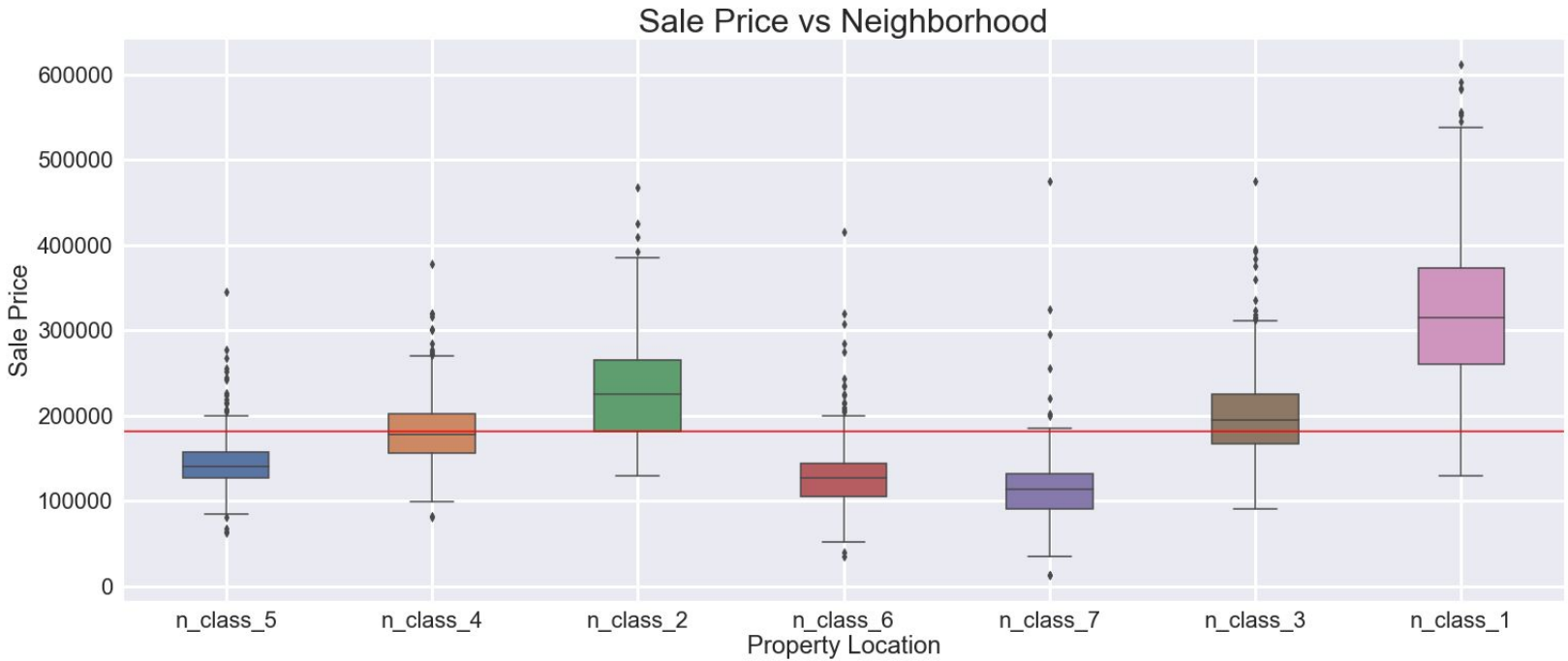
- Data matches up somewhat to a normal distribution with observations trailing off at higher sale prices and a high number of observations in the first standard deviation below the mean.
- Very few outliers, but as the sales price increase variance increase (more interaction terms needed in future)
- Certain features line up very well with large developments while others do provide some slices to help identify variation from mean
- Feature engineering can be helpful but must be applied in distinct slices
    - This opens up opportunity in the future for new data to be brought in to better describe the properties


Sale Price Histogram with Sample Mean: 181470
Count
0
50
100
150
200
0
100000
200000
300000
400000
500000
600000
Sale Price


Sale Price vs Ground Level Living Space
Sale Price
600000
500000
400000
300000
200000
100000
0
1000
2000
3000
4000
5000
Ground Level (ft^2)

Pitfalls of Feature Engineering

Neighborhood Categories



Model Performance

- After cleaning and processing data was fit to a linear regression model and compared to a baseline for reference
  - Null model R^2 Training: -4.33*10^-4
  - Null model R^2 Test: -1.56*10^-10
  - Lasso R^2: 0.873
  - Lasso RMSE: 2.80*10^4
- Believe the model suffers from overfitting at this point due to the lack change when adding more features (or more precise interaction terms are needed for better performance)
- While it was possible to get slightly higher metrics with certain features, cross split scores did not improve indicating overfitting issues

Conclusions and Future Studies

- Model is good at describing sale price but could be improved to give better insight on edge cases
- Current model is good at predicting pricing has location, exterior features, accounts for garage and basement aspects
- More data on the exterior elements of the properties
- More extensive feature engineering
    - Specific interactions
    - Squared terms