

# Housing Data Analysis

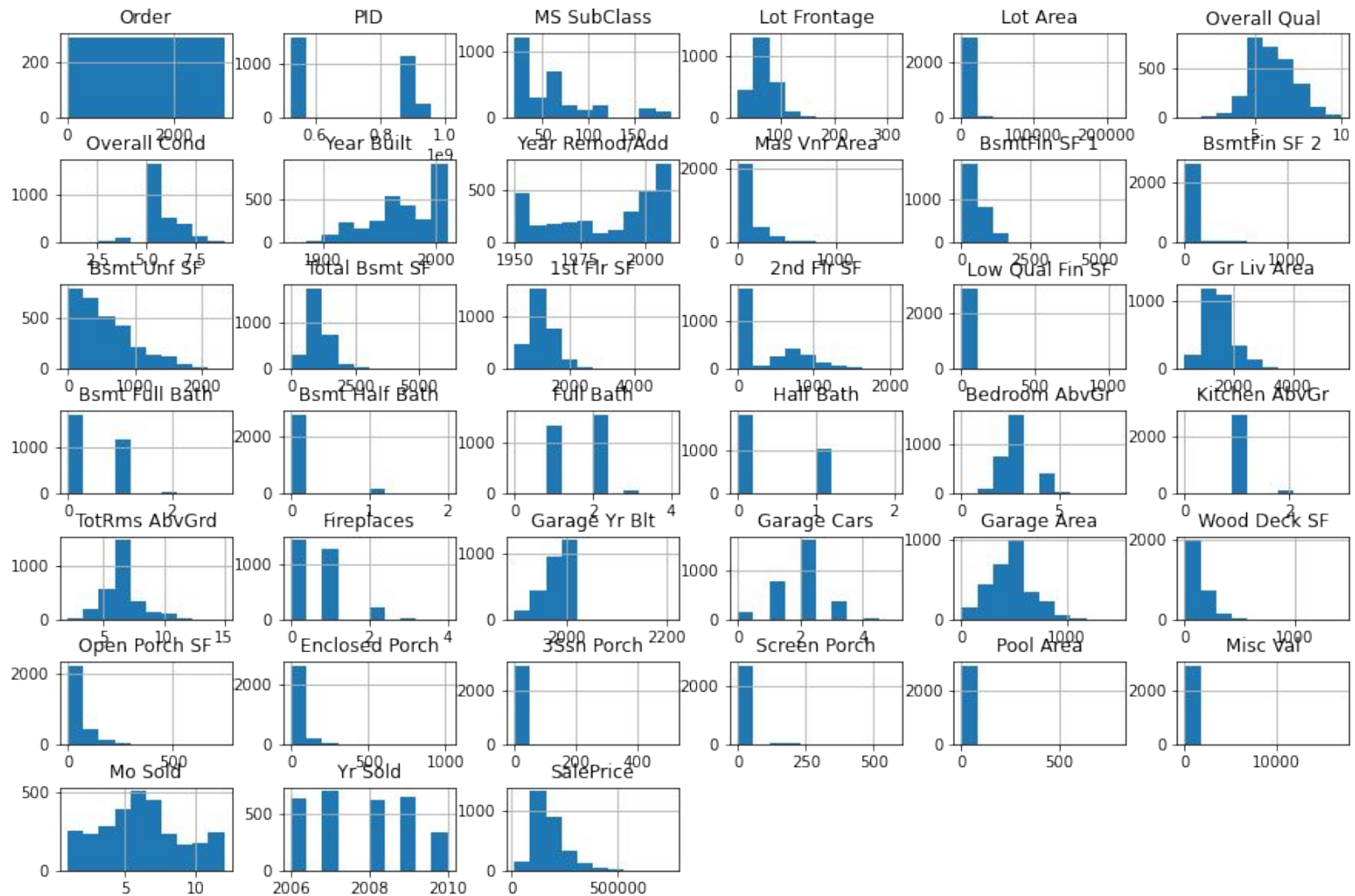
Ames Iowa

Springboard - Second Capstone

Andrew Seal

# Dataset

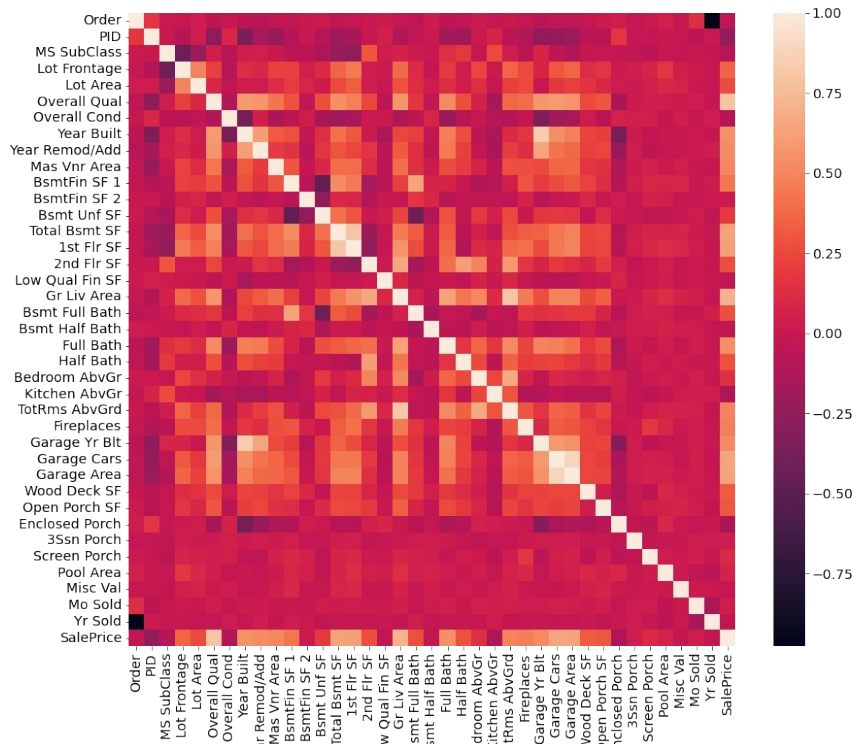
- Publicly available dataset with 2,930 unique sales between 2006 and 2010.
- 82 columns each representing a feature of the home.
- Includes categorical data:
  - Zoning
  - Neighborhood
  - Condition of the home...etc
- And numerical data:
  - Square footage
  - Lot size
  - Number of Bedrooms...etc



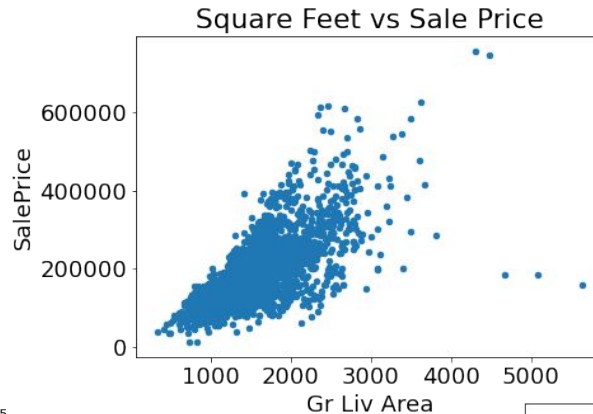
# Data Cleaning and Data Wrangling

- Check for Duplicates (Parcel ID)
- Visually inspect histograms for outliers or errors (previous slide)
- Columns with mostly missing data
- Columns with mostly homogeneous data

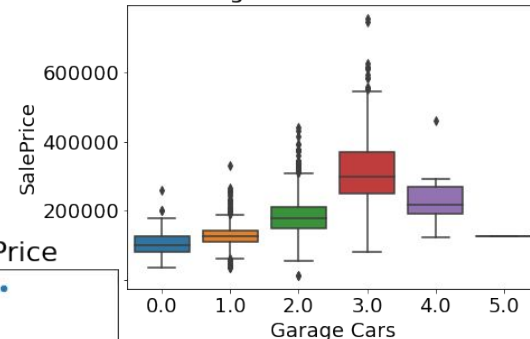
# Exploratory Data Analysis



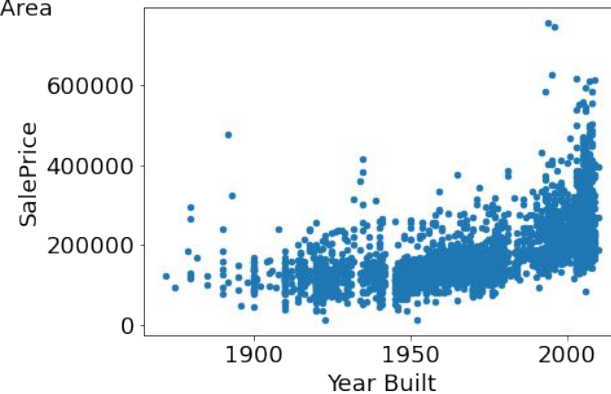
Square Feet vs Sale Price



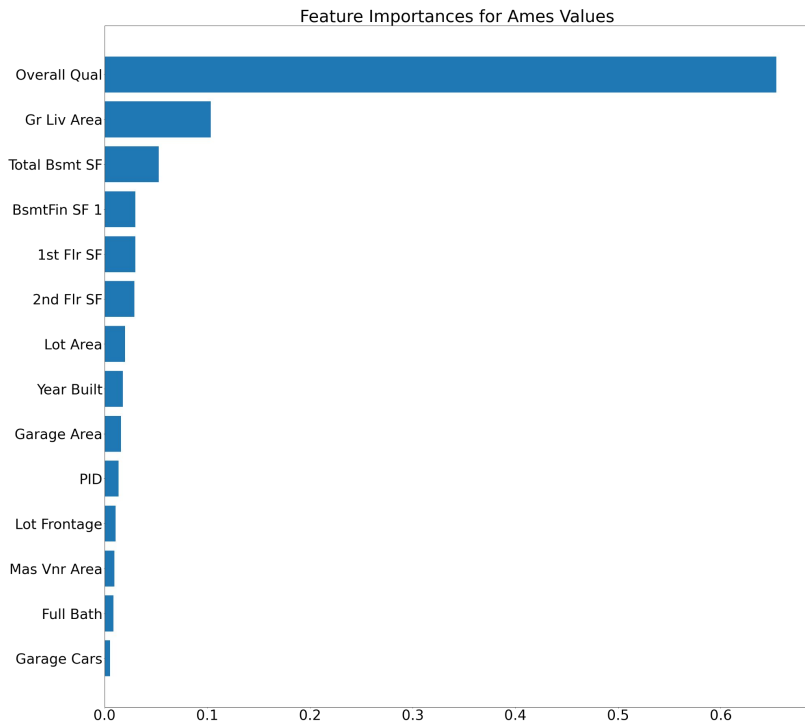
Garage Cars vs Sales Price



Year Built vs Sale Price



# Exploratory Data Analysis Continued



Random Forest Feature Importance

Overall Qual had outsized importance

Lack of information about what this variable is

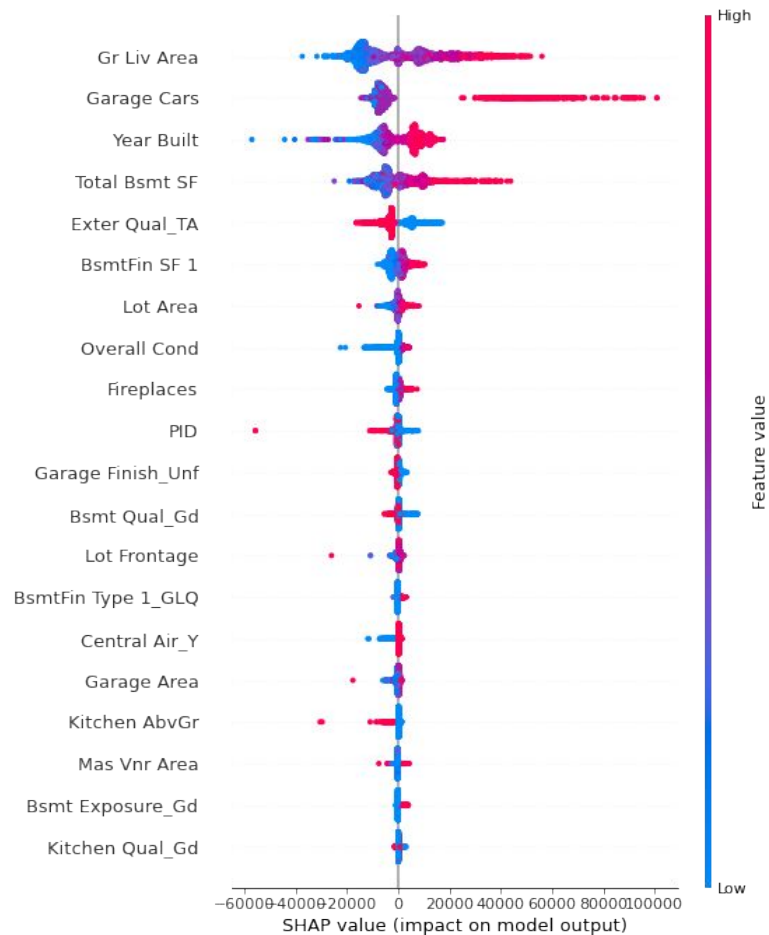
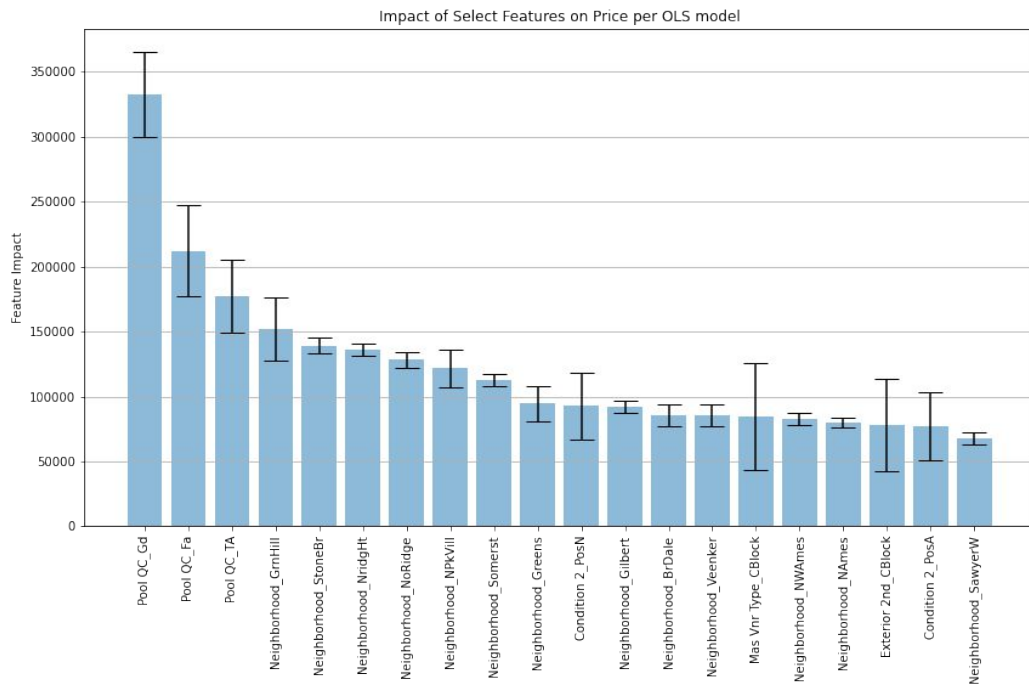
Ultimately Removed this Variable

# Preprocessing and Training

- Imputing Values:
  - Imputed MEAN for Lot Frontage (490 sales were missing this data 16.7%)
  - Imputed MODE Masonry Veneer Type and Masonry Veneer Area (23 sales)
  - Imputed MODE for Electrical (1 sale)
  - DID NOT Impute data for Garage, Air Conditioning, 2<sup>nd</sup> floor and other features that may not be present in each home
- Create Dummy Variables
  - K-1 Variables to avoid unnecessary collinearity
- NOW Drop Missing Data – Results in 2,747 Rows and 244 Columns
- Address Multicollinearity – VIF Analysis AND Simple Collinearity
- Standardize data using Robust Scaler

# Feature Importance

1. Coefficients from OLS Regression
2. SHAP Feature Importance (right)



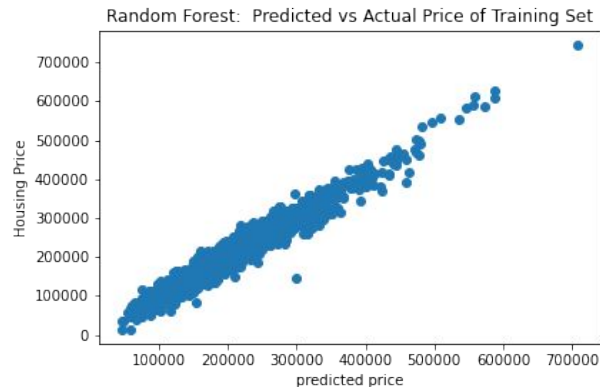


# Train Test Split

- Training Set with 75% of the Sales
- Test Set with remaining 25% of the Data

# Random Forest Model

- Optimized Hyperparameters:
  - N\_estimators
  - Max\_depth
- Used Random Search for CV



## Random Forest Evaluation Metrics

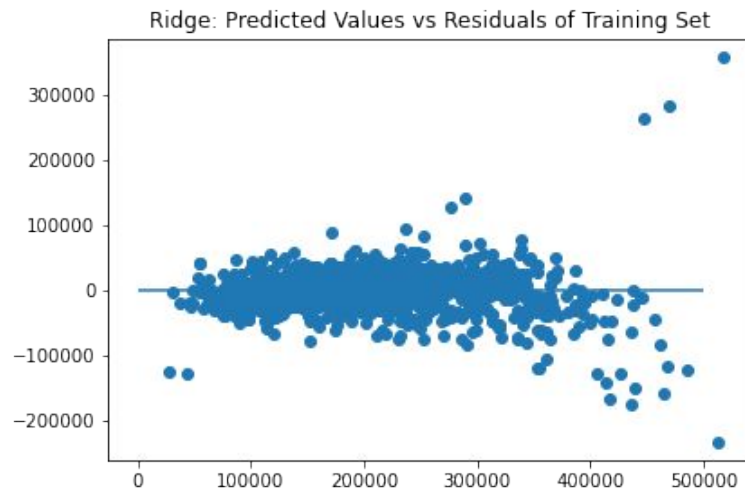
## Model: All Features

## Model: Top Twenty Features

r2 Score Train	0.983	0.983
r2 Score Test	0.862	0.856
Mean Absolute Error (MAE) Train	6189.75	6308.99
Mean Absolute Error(MAE) Test	17905.1	18336.8
Root Mean Square Error (RMSE) Train	10268.8	10250.1
Root Mean Square Error (RMSE) Test	29797.4	30411.4
Optimal n_estimator	5308	10000
Optimal max_depth	20	110

# Ridge Regression

- Useful when multicollinearity exists.
- Grid Search CV to optimize Alpha



Ridge Regression Evaluation Metrics

Model: All Features

Model: Top Twenty Features

r2 Score Train

0.88

0.789

r2 Score Test

0.883

0.829

Mean Absolute Error (MAE) Train

17309.2

22532

Mean Absolute Error(MAE) Test

18166.6

22350.1

Root Mean Square Error (RMSE) Train

27466.3

36415.1

Root Mean Square Error (RMSE) Test

27442.7

33171.5

Optimal Alpha

6

35

# Gradient Boosting

- Grid Search Cv
- To optimize Learning Rate

Gradient Boosting Evaluation Metrics	Model: All Features	Model: Top Twenty Features
r2 Score Train	0.974	0.97
r2 Score Test	0.89	0.881
Mean Absolute Error (MAE) Train	9603.46	10340.5
Mean Absolute Error(MAE) Test	16981	17789.9
Root Mean Square Error (RMSE) Train	12869.2	13781.8
Root Mean Square Error (RMSE) Test	26604.8	27666.8
Optimal Learning Rate	0.1	0.1
Optimal Max Depth	4	4

# Comparing The Models

- Random Forest performs best on Training Data
- Gradient Boosting has highest R-squared and Lowest RMSE on test data
- Expect Gradient Boosting to most accurately predict Sales Price

Model	r2 Score (Test-Data)	Root Mean Square Error (Test Data)	Hyperparameter 1	Hyperparameter 2
Random Forest (All Features)	0.862	29797.42	n_estimators = 5,308	max_depth = 20
Random Forest (Top 20 Features)	0.856	30411.44	n_estimators = 10,000	max_depth = 110
Ridge Regression (All Features)	0.883	27442.69	Alpha = 6	
Ridge Regression (Top 20 Features)	0.829	33171.55	Alpha = 35	
Gradient Boosting (All Features)	0.89	26604.81	Learning Rate = 0.1	max_depth = 4
Gradient Boosting (Top 20 Features)	0.881	27666.77	Learning Rate = 0.1	max_depth = 4

# Gradient Boosting

- Selected as best model

Selected Model		Gradient Boosting (All Features)	
-----		-----	
Model		Gradient Boosting (All Features)	
Hyperparameter 1		Max Depth = 4	
Hyperparameter 2		Learning Rate = 0.1	
r2 Score Train		0.974	
r2 Score Test		0.89	
Root Mean Square Error (RMSE)	Train	12869.0	
Root Mean Square Error (RMSE)	Test	26605.0	

# Test Case

- Randomly Selected ONE property from our Test Data
- Applied Gradient Boosting Model (ALL Features) to predict Sales Price
- Model predicted Sales Price of \$247,968
- Actual Sales Price of \$226,000

# Conclusion and Next Steps

Three models evaluated

- Random Forest, Ridge Regression, Gradient Boosting
- For all models looked at two feature sets

Compared models on R-Squared and RMSE to select best model

Best model was Gradient Boosting (All Features)

Next Steps:

- Further optimization of hyperparameters
- Additional feature sets
- Assess application for other real estate types (commercial , multifamily)