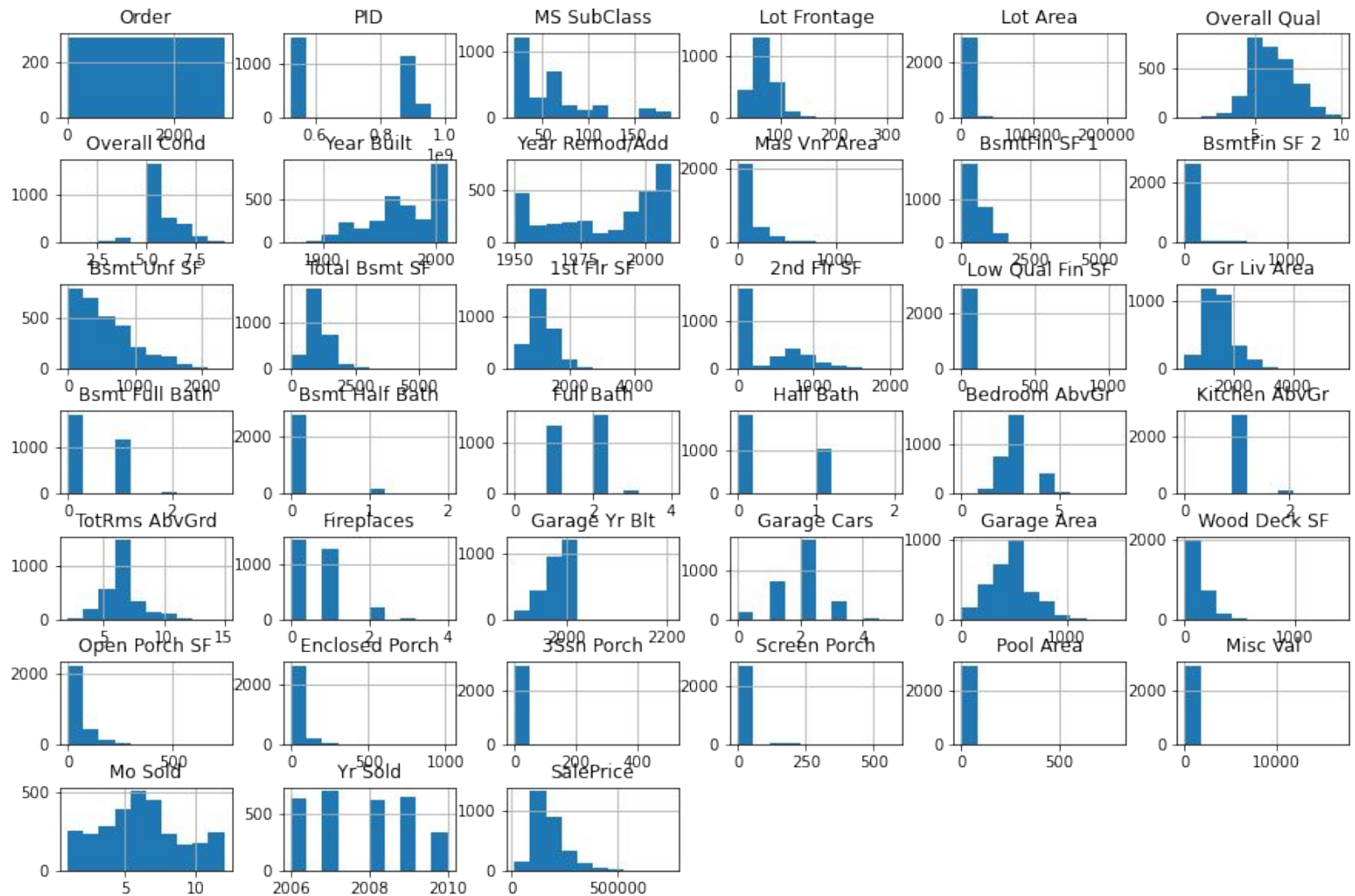# Housing Data Analysis

Ames Iowa

Springboard - Second Capstone
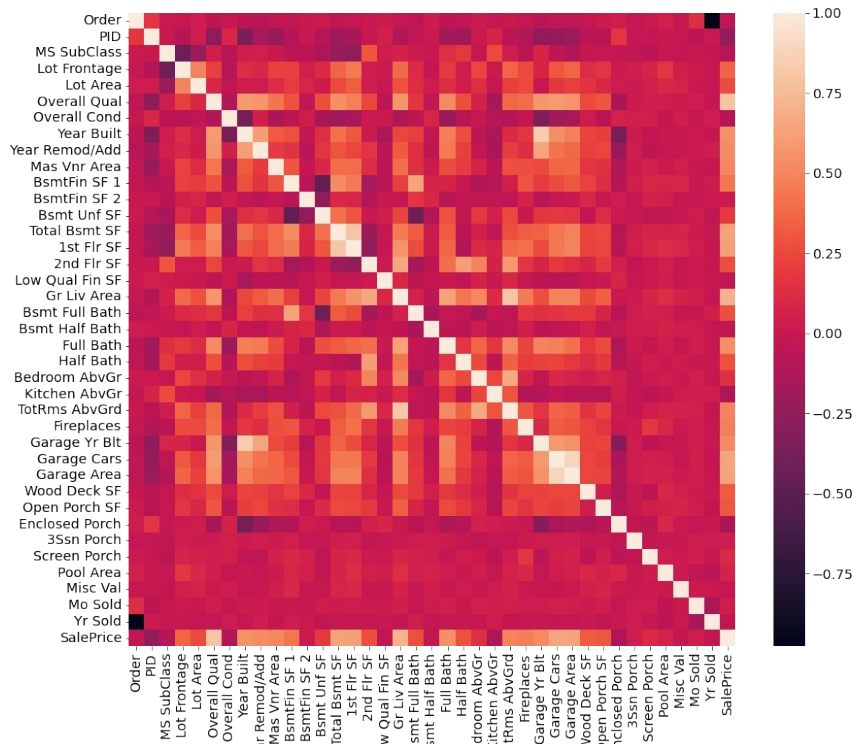
Andrew Seal

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
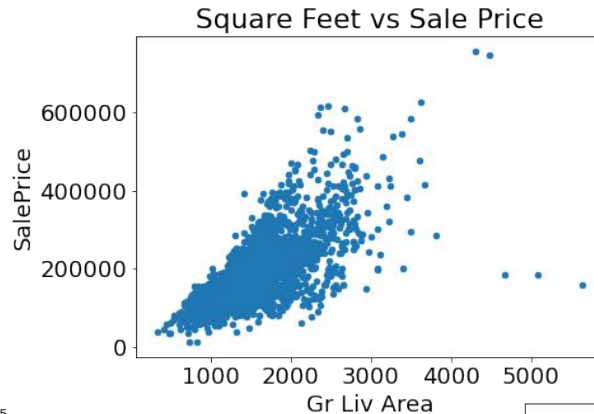 - Number of Bedrooms…etc

Data Cleaning and Data Wrangling

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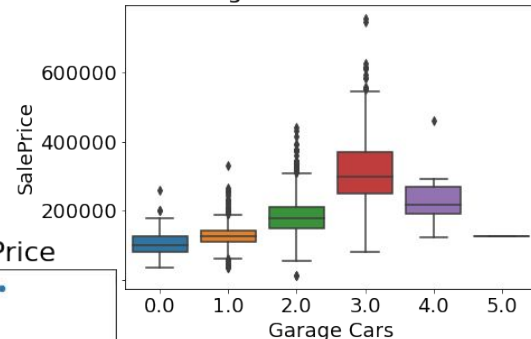- Columns with mostly homogeneous data

Exploratory Data Analysis

Garage Cars vs Sales Price

Square Feet vs Sale Price

Year Built vs Sale Price
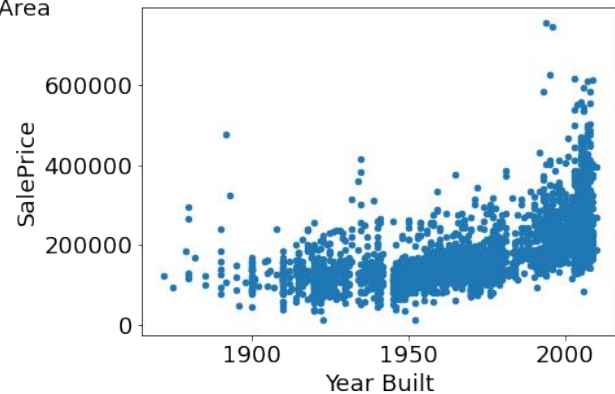
Exploratory Data Analysis Continued

Feature Importances for Ames Values
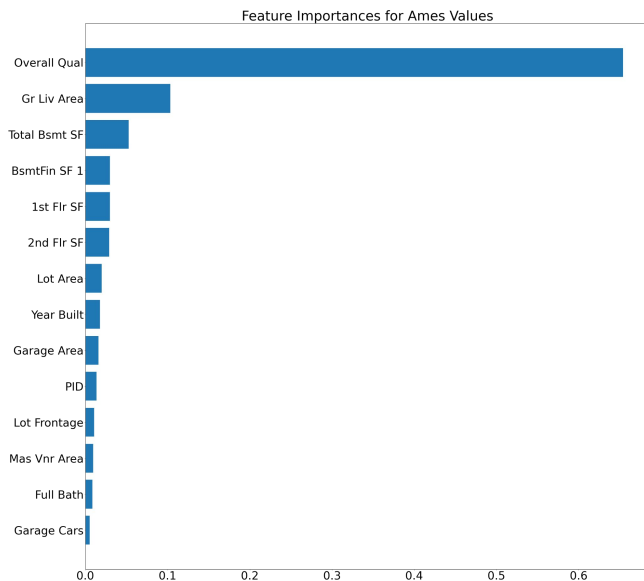
- Random Forest Feature Importance
- Overall Qual had outsized importance
- Lack of information about what this variable is
- Ultimately Removed this Variable

Select Feature Importances for Ames Values EXCLUDING Overall Qual
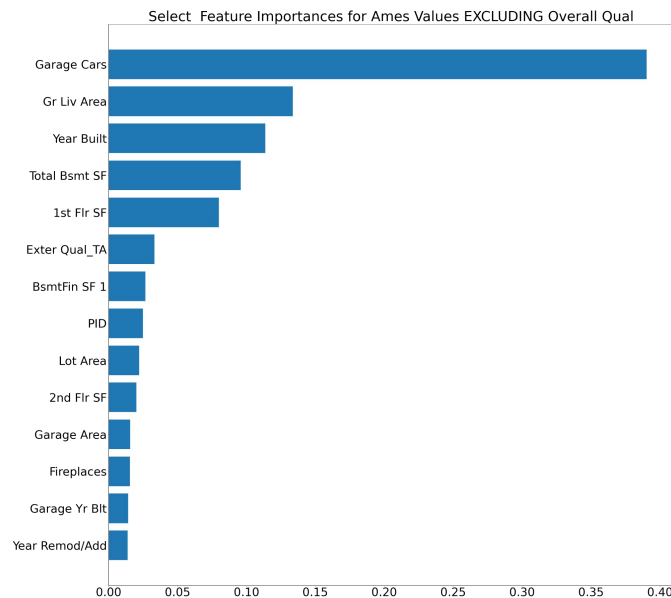
Preprocessing and Training

- Imputing Values:
  - Imputed MEAN for Lot Frontage (490 sales were missing this data 16.7%)
  - DID NOT Impute data for Garage, Air Conditioning, 2nd 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

Feature Importance - OLS (Unscaled)

- Unscaled coefficients should correspond to $ impact on sales price.
- While a pool may detract from sales price, the magnitude of the impact is unreasonable.

```
coefs.sort_values(ascending=False).head(5)
```

```
Neighborhood_GrnHill      136051.193859
Condition_2_PosA           90250.131444
Electrical_Mix             70502.849164
Neighborhood_StoneBr       68781.758793
Neighborhood_NridgHt       57973.372031
dtype: float64
```

```
coefs.sort_values(ascending=False).tail(5)
```

```
Condition_2_PosN          -73622.193755
Mas_Vnr_Type_CBlock       -91966.406850
Pool_QC_TA               -158684.213611
Pool_QC_Fa               -188326.499476
Pool_QC_Gd               -331994.124866
dtype: float64
```

Feature Importance

1. Coefficients from OLS Regression (standardized using Robust Scaler)
2. SHAP Feature Importance (right)
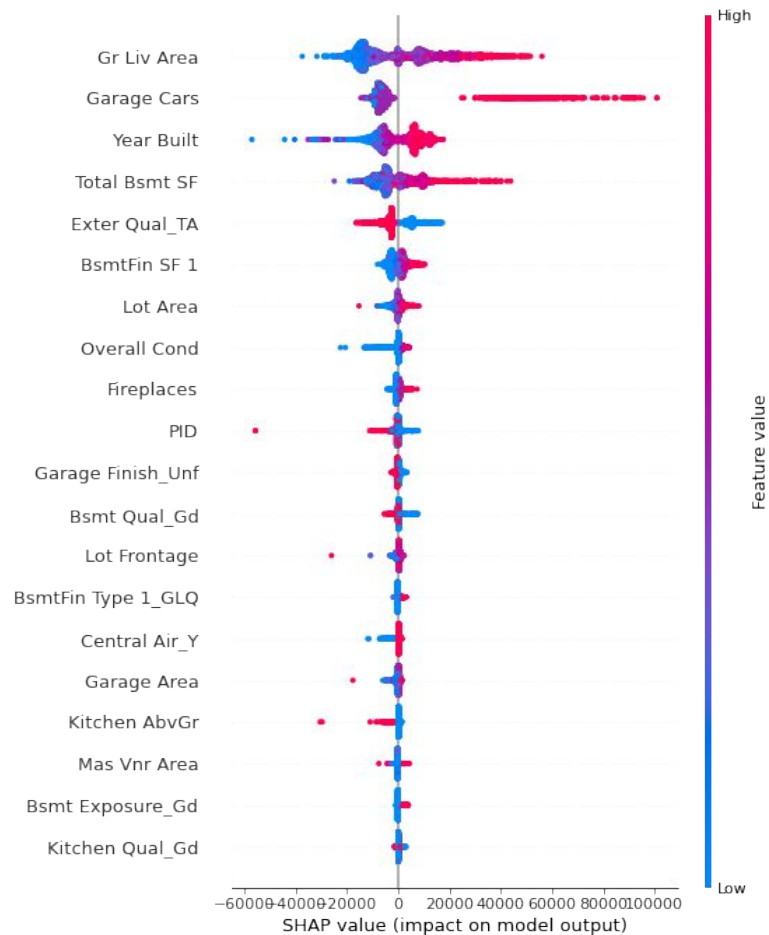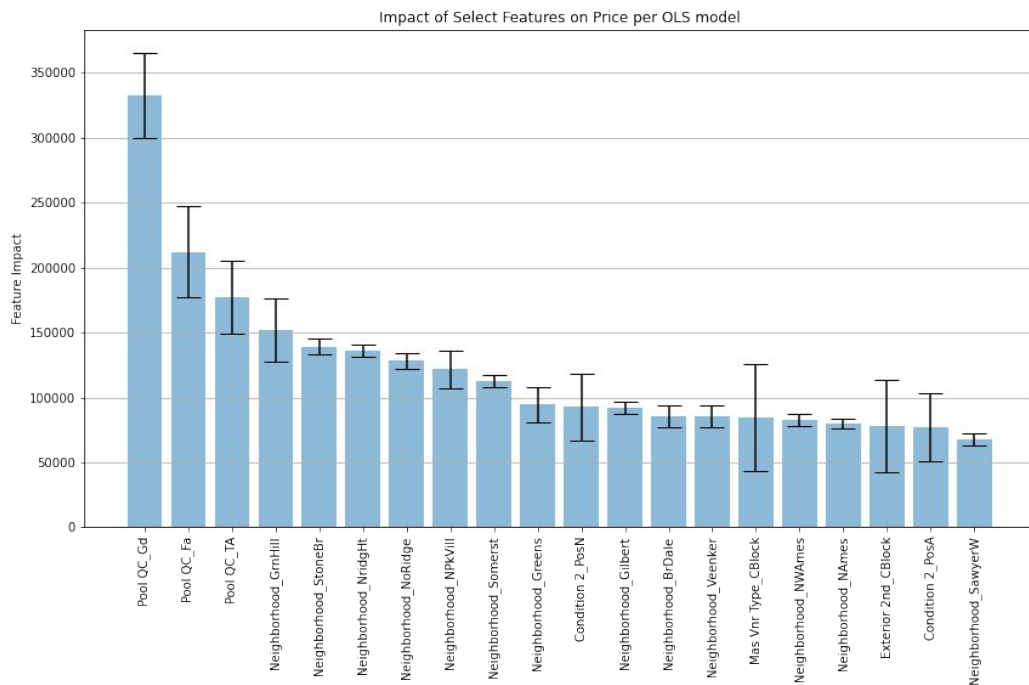
Train Test Split

- 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: Predicted vs Actual Price of Training Set
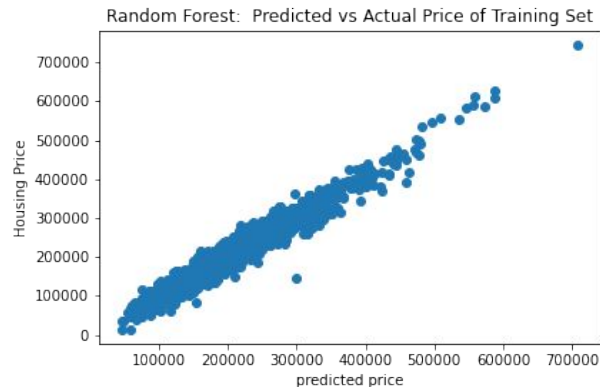
Random Forest Evaluation Metrics	Model: All Features	Model: Top Twenty Features
r2 Score Train	0.983	0.983
r2 Score Test	0.862	0.863
Mean Absolute Error (MAE) Train	6180.41	6329.58
Mean Absolute Error(MAE) Test	17881.2	18017.3
Root Mean Square Error (RMSE) Train	10277.5	10454.4
Root Mean Square Error (RMSE) Test	29823	29733
Optimal n_estimtor	5308	10000
Optimal max_depth	20	110

Ridge Regression

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

Ridge: Predicted Values vs Residuals of Training Set
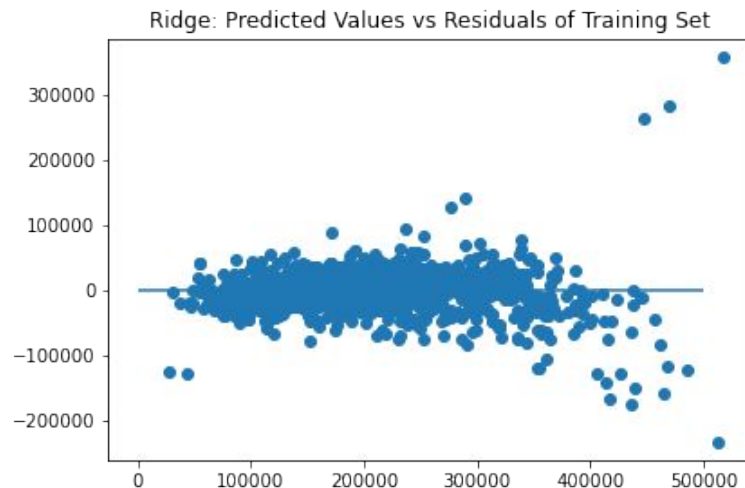
Ridge Regression Evaluation Metrics	Model: All Features	Model: Top Twenty Features
r2 Score Train	0.88	0.789
r2 Score Test	0.883	0.829
Mean Absolute Error (MAE) Train	17309.2	22532
Mean Absolute Error(MAE) Test	18166.6	22350.1
Root Mean Square Error (RMSE) Train	27466.3	36415.1
Root Mean Square Error (RMSE) Test	27442.7	33171.5
Optimal Alpha	6	35

Gradient Boosting

- Grid Search Cv
- To optimize Learning Rate and Max Depth

```
Gradient Boosting Evaluation Metrics      Model: All Features      Model: Top Twenty Features
----------------------------------        --------------------     ----------------------------
r2 Score Train                                      0.974                          0.983
r2 Score Test                                       0.889                          0.886
Mean Absolute Error (MAE) Train                    9603.46                        7756.69
Mean Absolute Error(MAE) Test                        16994                        17109.9
Root Mean Square Error (RMSE) Train                12869.2                          10202
Root Mean Square Error (RMSE) Test                 26719.8                        27106.1
Optimal Learning Rate                                 0.1                            0.1
Optimal Max Depth                                       4                              5
```

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 precisely predict Sales Price

Model	r2 Score (Test-Data)	Root Mean Square Error (Test Data)	Hyperparameter 1	Hyperparameter 2
Random Forest (All Features)	0.862	29823.01	n_estimators = 5,308	max_depth = 20
Random Forest (Top 20 Features)	0.863	29733.0	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)	0.889	26719.8	Learning Rate = 0.1	max_depth = 4
Gradient Boosting (Top 20 Features)	0.886	27106.08	Learning Rate = 0.1	max_depth = 5

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                            0.889
Root Mean Square Error (RMSE) Train      12869.0
Root Mean Square Error (RMSE) Test       26720.0
```

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)