Project 2:
Zillow Predictive Modeling

Linear Regression Model for Housing Prices in Ames, IA
General Assembly DSI-NY-6

Thomas Ludlow
December 7, 2018

Housing Prices in **Ames, IA:**

What features most affect sales prices?

What modeling approaches yield the most accurate predictions?

Target: **Sales Price**

Source: https://www.kaggle.com/c/dsi-us-6-project-2-regression-challenge/data

Source Data

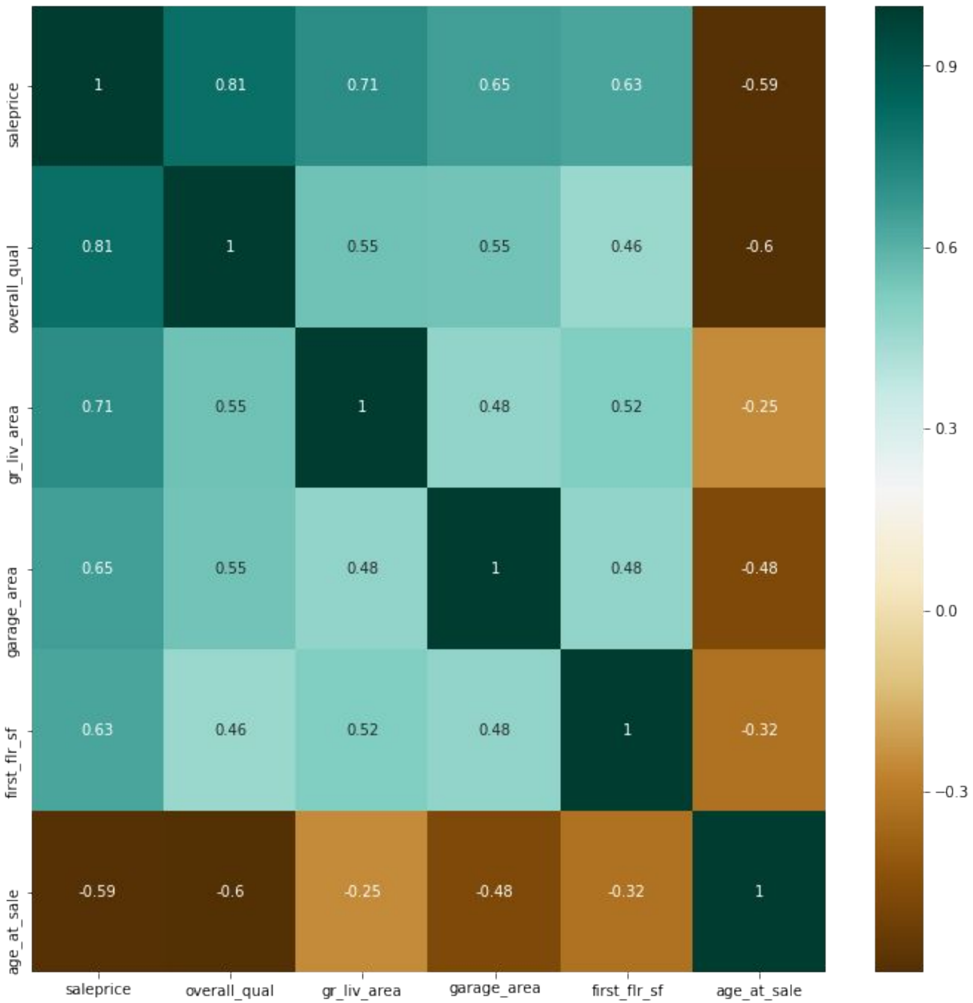
- Records from **2,050** home/building sales in **Ames, IA** from **2006 - 2010**
- **80** pieces of building details including:
  - Years of construction, sale, and remodel
  - Neighborhood, proximity to transportation/parks & recreation
  - Building type and municipal subclass
  - Building materials for exterior, roofing, masonry
  - Number of rooms, area in sq. ft.
  - Utility details
  - Lot details such as size, shape, incline
  - Details on sale execution
  - Quality and condition ratings

Exploratory Data Analysis (EDA)

- Look at data for completeness
  - Fix missing data if possible
  - Remove corrupted rows
- Reshape for usability / accuracy
  - Convert years into ages
  - Turn Y/N into 1/0
  - Standardize category names/spellings
- Identify outliers
  - Determine whether to keep or remove, and for which categories

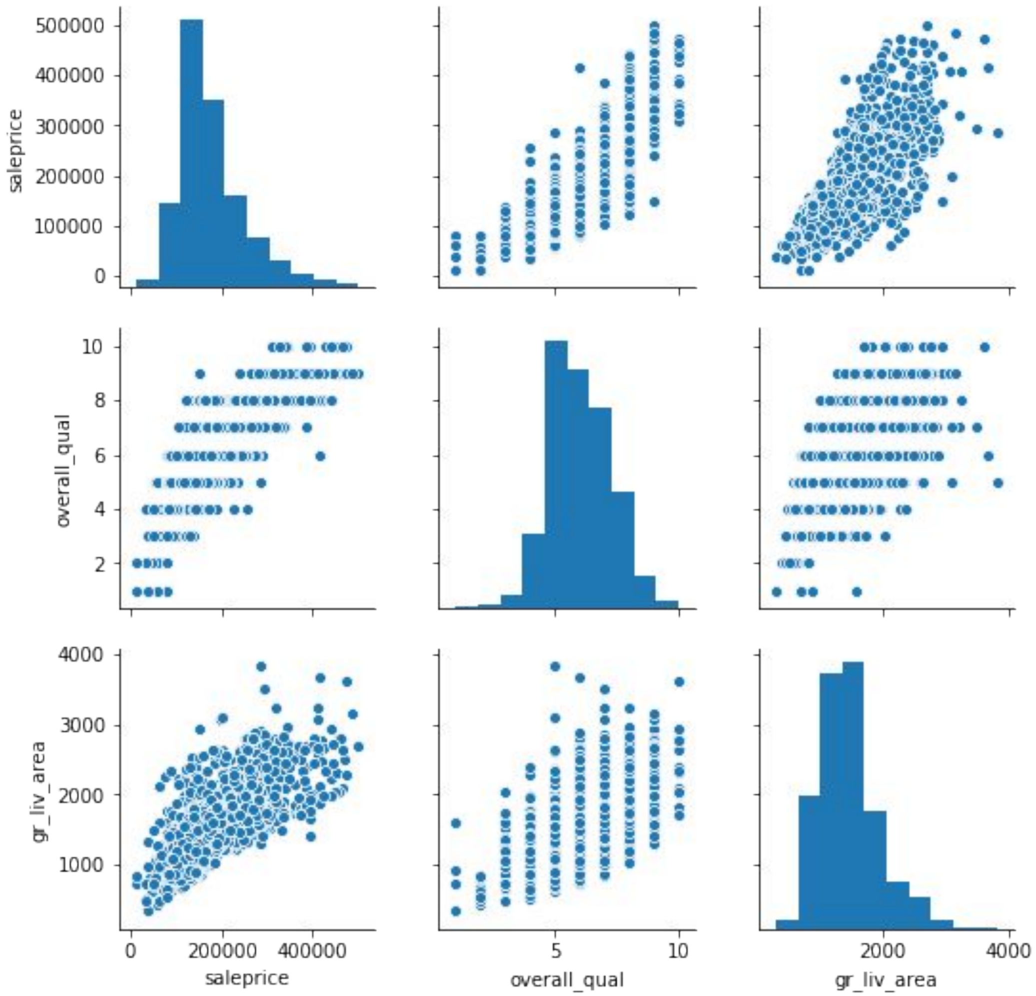
Feature Exploration

Heat mapping helps us visualize correlations between variables



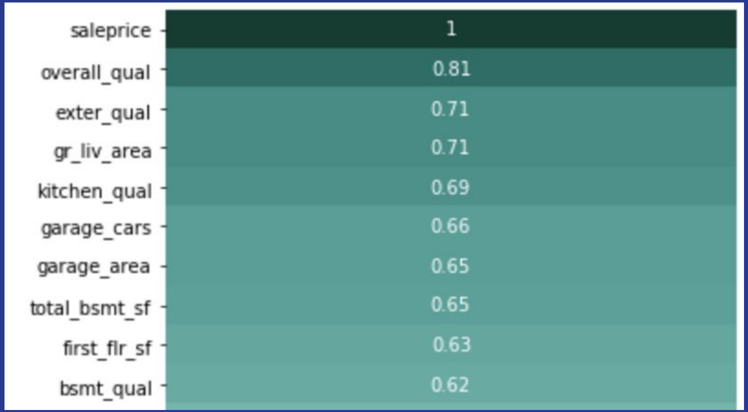
Feature Exploration

Pair plotting shows us distribution relationships between variables

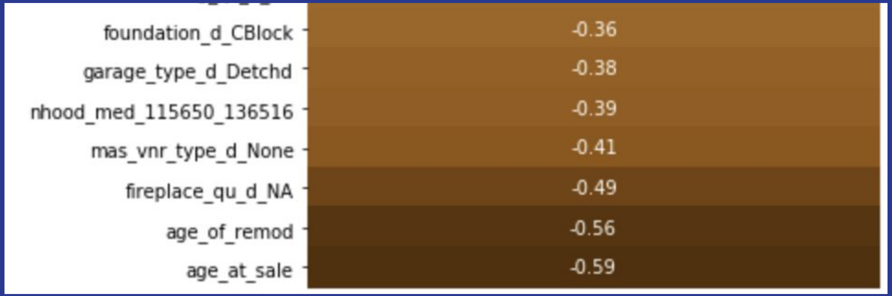


Which features most affect sales prices?

	Feature	Corr.
1.	Overall Quality	(0.81)
2.	Exterior Quality	(0.71)
3.	Above-Grade Living Area	(0.71)
4.	Kitchen Quality	(0.69)
5.	Garage Number of Cars	(0.66)
6.	Garage Area	(0.65)
7.	Total Basement Square Ft	(0.65)
8.	First Floor Square Ft	(0.63)
9.	Basement Quality	(0.62)
10.	Age at Sale	(-0.59)



High positive and negative correlations

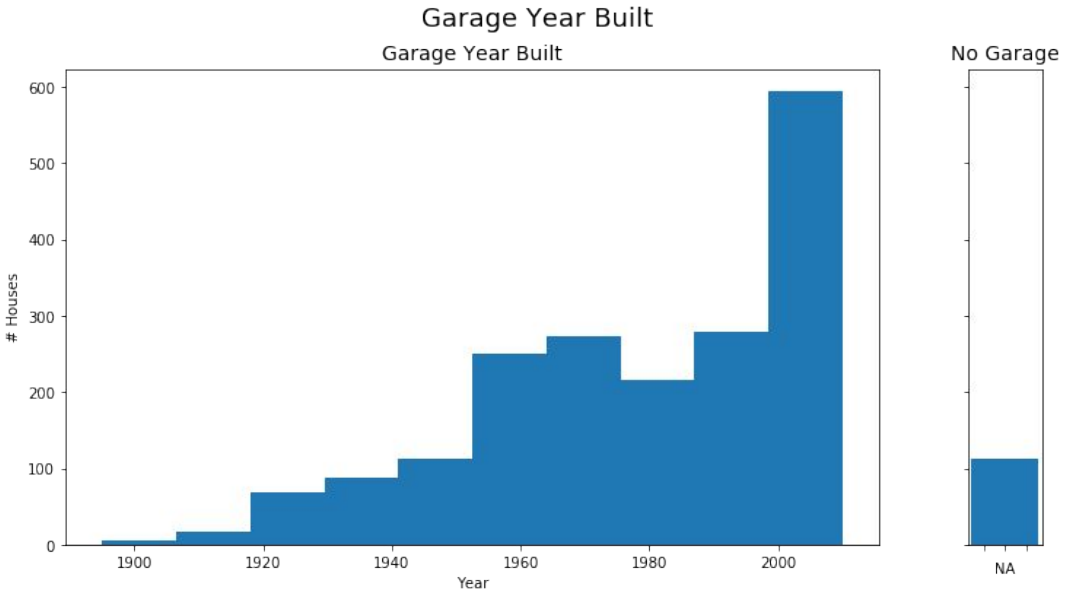


What modeling approaches yield the most accurate predictions?

We used a ***Linear Regression Model*** to test.

Modeling Techniques: Dummy Variables

- Convert variables to **dummies**



- Garage has significant impact, but **112** houses had no garage
- Each value gets its own binary 1/0 variable
- Using dummy variable "has_garage" retains important detail
- For "ms_subclass", dummies can represent shared qualities (*e.g., "is_post_war", "is_two_story", etc.)*

Modeling Techniques: Ordinal Values

- Assign **ordinal values** to “quality” and other scaled variables

- Change from text categories into numerical values for model processing
- “Ex”, “Gd”, “Av”, “Fr”, “Po” become 5, 4, 3, 2, 1
- Mapped in addition to creating dummy categories, and estimated values based on materials

Modeling Techniques: Polynomials

- Linear regression models can be enhanced with **interactions** and **polynomial** variables

- Added interaction and polynomial variables for the top-14 correlated categories against all categories
- Added two cubic values by top-14 correlations
    - Overall_qual Gr_liv_area
    - Exter_qual Gr_liv_area

Model Execution

- Train / Test Split
  - **Separate portion of data** as control group to assess model performance before submitting
- Scaling
  - Using tools in Python library SciKit-Learn, convert numeric values to **standard deviation** of all values for a variable
- Power Transform
  - Tools convert variables **logarithmically** to even out skewed distributions

Model Execution

- Regularization
  - While modeling, we used two types of regularizing models: Ridge and Lasso
  - Models impose costs on using too many variables to improve accuracy
  - Lasso reduces predictive coefficients to zero quickly, helping to identify unneeded variables
- Assessment
  - Test models using Cross-Validation Scoring to determine R^2 score
  - Comparing R^2 scores of training and test data tells modeler about fit and predictive accuracy

What values did the model find most important?

Final Ridge model used **111** variables.

The variables with the largest coefficients had the biggest impact on the model's predictions.

Top 10 Model Coefficients

1. Total Basement SqFt x Basement Quality
2. Overall Quality^2
3. Above Grade Living Area x Fireplace Quality
4. Above Grade Living Area x Basement Quality
5. Kitchen Quality x Garage Area
6. Overall Quality x Exterior Quality
7. Overall Quality x Kitchen Quality
8. Kitchen Quality x Garage Cars
9. Overall Quality x Garage Area
10. Overall Condition

What does this mean for Zillow?

- The Ames, IA model can serve as a starting point for similar markets

- This model can be further optimized using GridSearch hyperparameter techniques

- An iterative modeling approach will continue to improve predictive results in all markets

- Additional research should be done into quantifying materials categories

Final Model Details

Regularization: Ridge

Scaling: StandardScaler

Power: PowerTransformer on variables between 1 and 30 absolute skewness

Approach: Initially built “kitchen sink” model and used Lasso with manual alpha to manage Convergence Errors

Removed zero-coefficient values from LassoCV to create Ridge model

Cross-Validation R^2 Score

0.9107

R^2 Score on Test Prediction

0.9214

Engineered Features

- **Custom Garage Dummies by Percentile Group**
  Created a loop that when given a percent number, automatically broke all garages into categories based on age, and assigned descriptive title to DataFrame

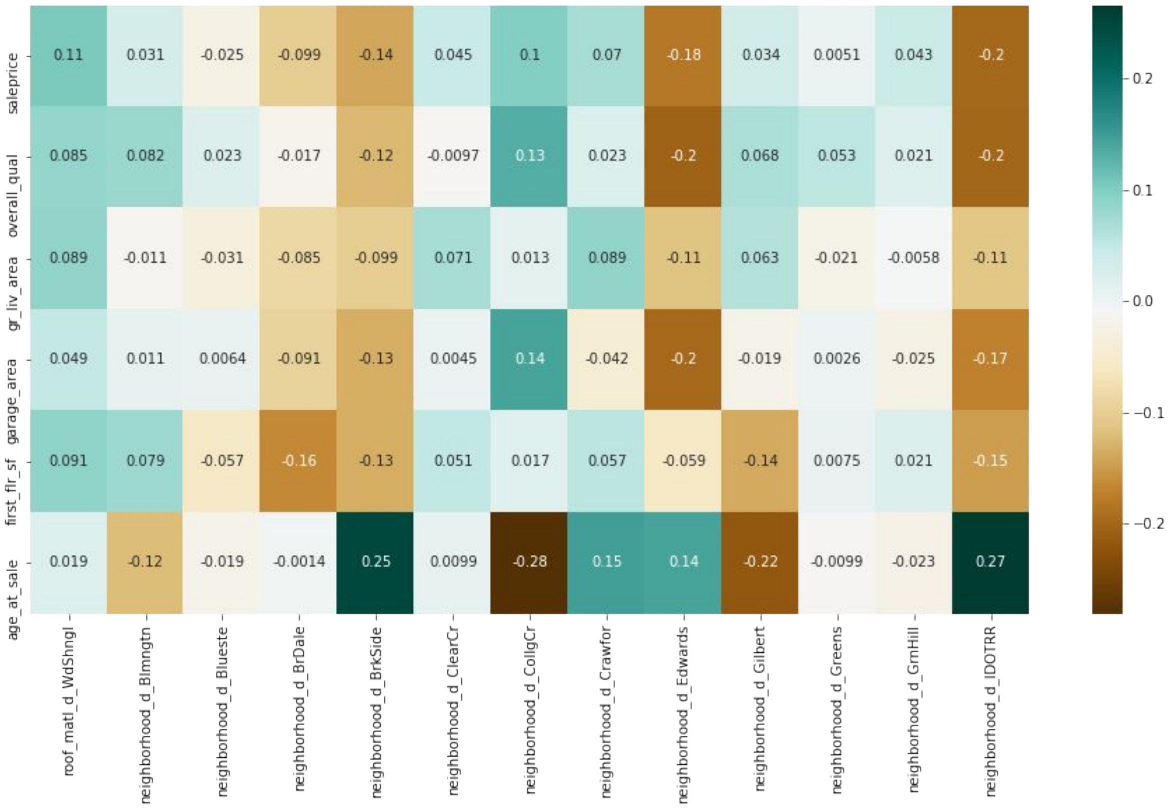
- **Custom Neighborhood Dummies by Percentile Group**
  Created a loop to group neighborhoods by median sale price by percentile, and assigned value ranges in DataFrame column names

- **Dictionary Ordinal Assignment**
  Used two dictionaries to allow adjustment of ordinal values and mass application to DataFrame

Engineered Features

- **Partial Heatmap Loop**
  Configured Seaborn heatmap to show variables correlation 12 at a time against 6 key variables

- **Partial PowerTransform**
  Created separate DataFrame to limit PowerTransform to necessary distributions *(resolved divide-by-zero errors)*



Questions?

Thank you!