

Ames, Iowa Housing Data

Source: Ames, Iowa Assessor's Office

2006 - 2010



Data Science Problem

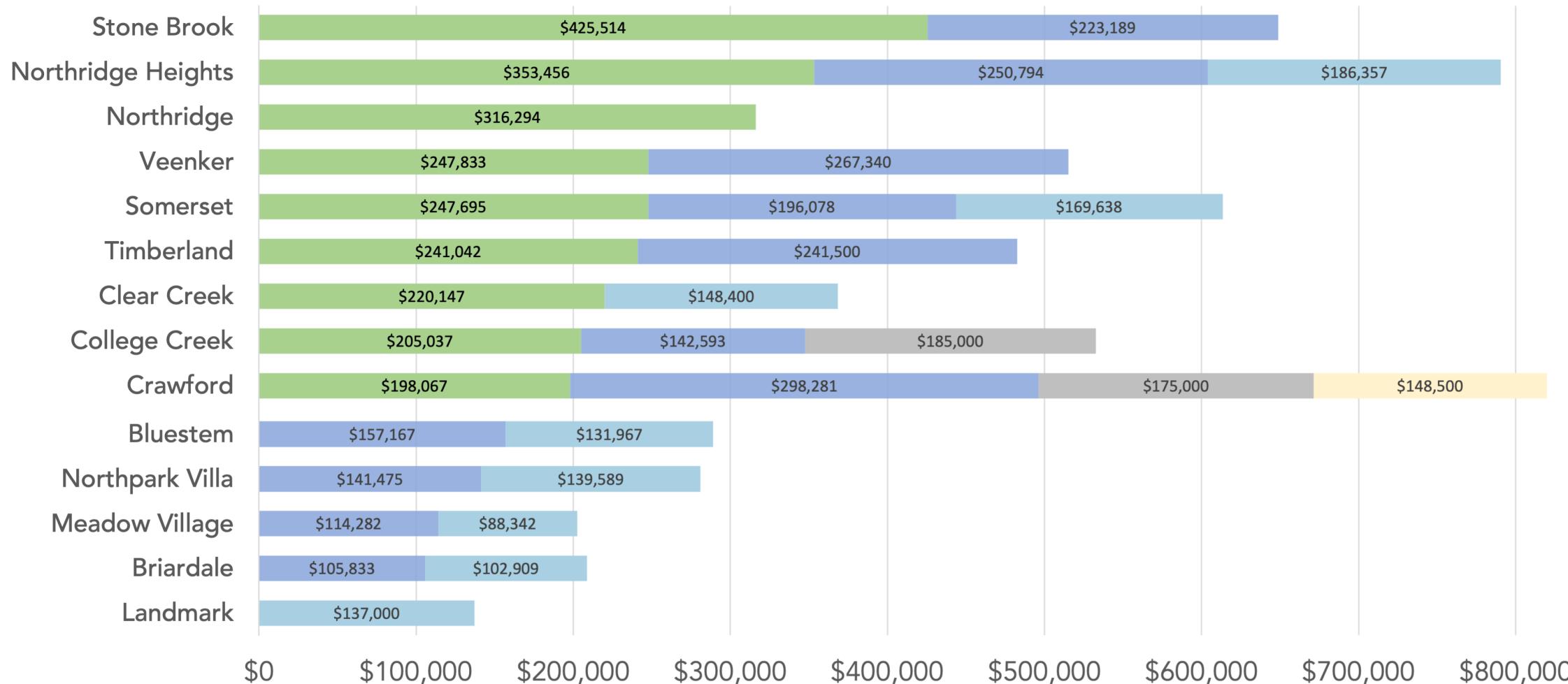
- ▶ Help the city of Ames, IA create the most predictive model possible given categorical and numerical data about various properties and their surrounding neighborhoods, to inform future data collection and pricing assessments.

Part I : Feature Selection

Find features that correlate with the target variable
(Sale Price) and understand how they interact with
target and other independent variables

Average Sale Price of a Home by Neighborhood

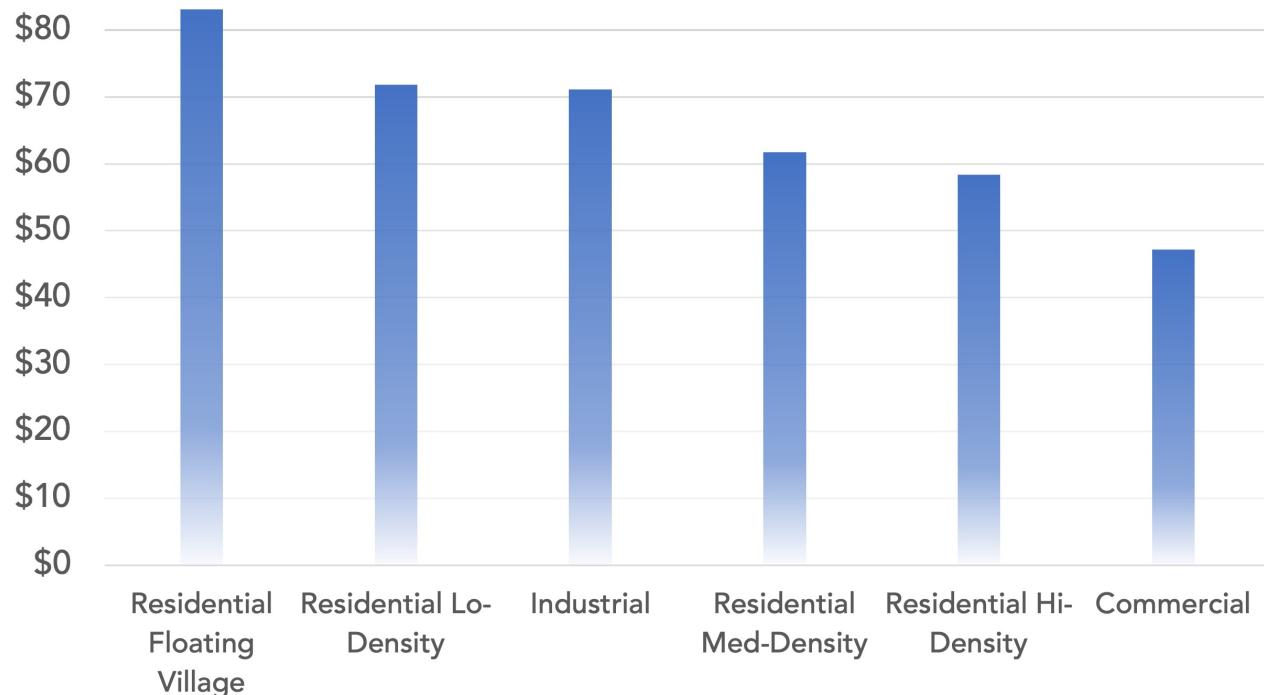
■ Single-Family Detach. ■ Townhouse End Un. ■ Townhouse Inside Un. ■ Duplex ■ Two-Family Conv.



Zoning

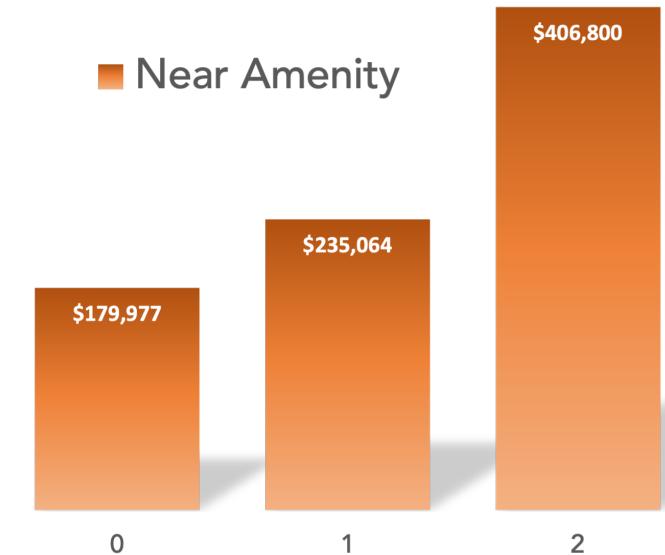
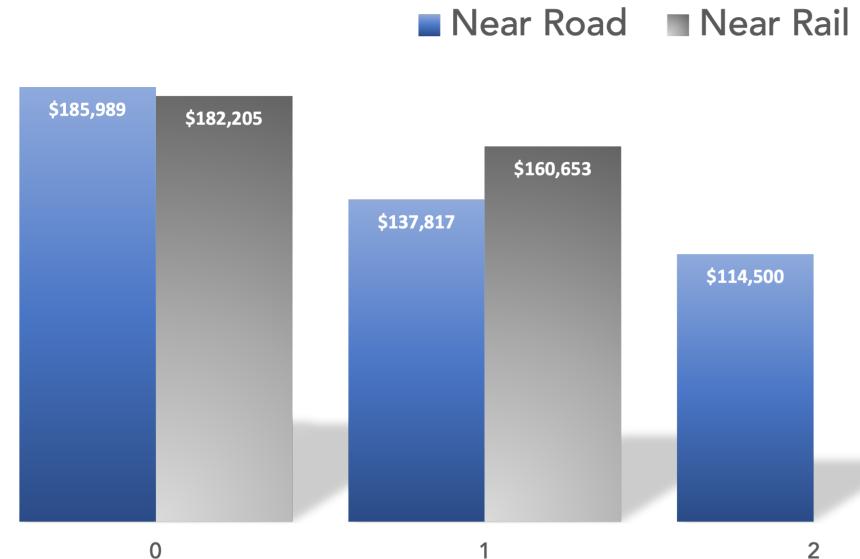
- ▶ Correlation does not equal causation
- ▶ Other variables could have contributed to the price ranges in those neighborhoods
- ▶ All Commercial properties located in the Iowa Dept. of Transportation and Railroad neighborhood

Price per Square Foot, by Zone

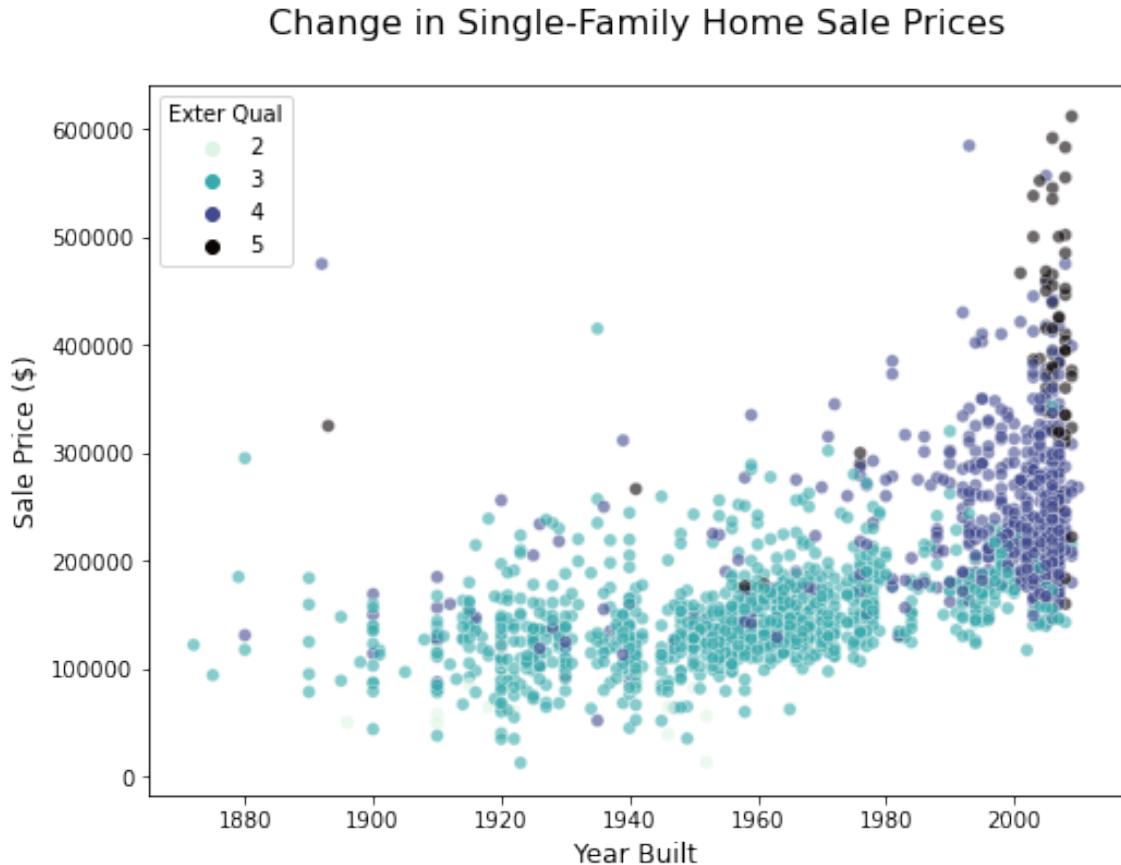


Proximity to Other Features

- ▶ Properties near positive features such as parks, outdoor recreation, etc. went up in value
- ▶ However, properties near roads or railroads went down in value



Qualitative Features



- ▶ There's been a greater emphasis on "curb appeal" in the last 10 years
- ▶ Lots of features on the quality, condition, and finishes of the property
- ▶ Important to rule out collinearity

Quantitative Features

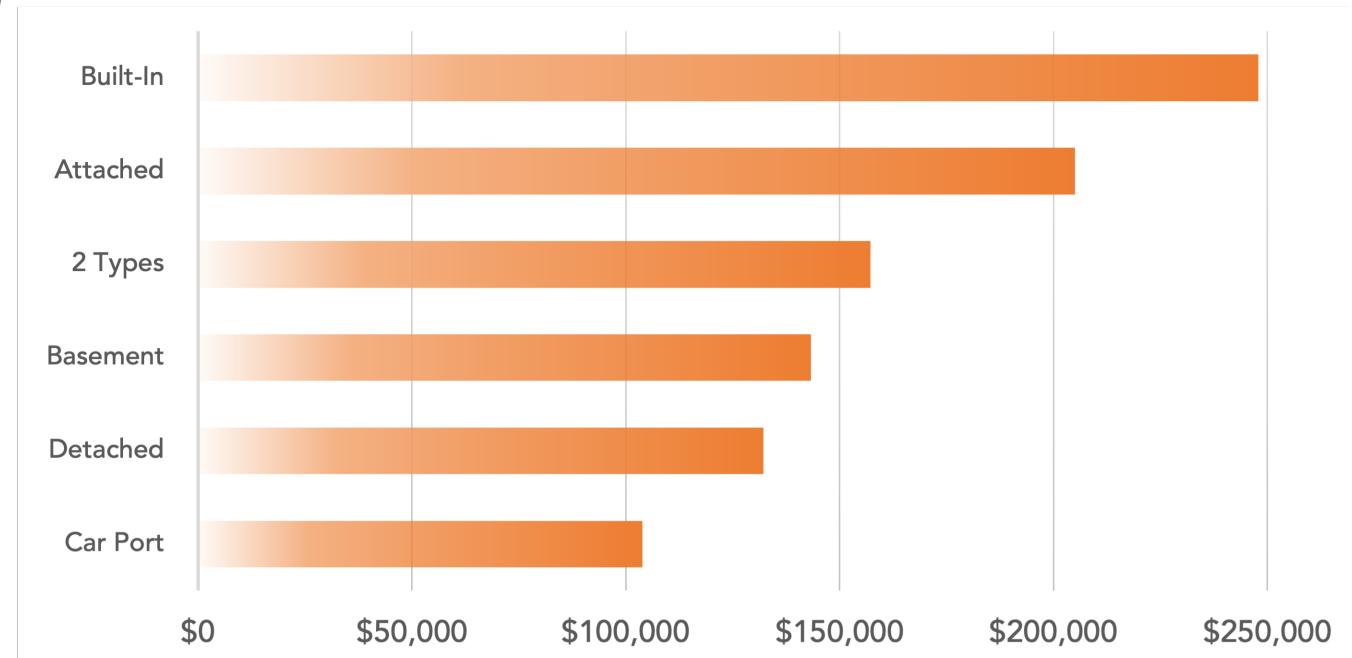


- ▶ Lots of variables pertaining to square footage of 1st Floor, 2nd Floor, and Basement areas
- ▶ The usual number of bedrooms and bathrooms
- ▶ Important to rule out duplicates here too

Miscellaneous

- Garages
- Sheds
- Pools
- Porches
- Decks

Sale Price by Garage Type



Part II : Feature Engineering

Standardize ratings, normalize scales, and encode categorical information among selected features

Categorical Rankings



Neighborhood (1-28)

- Least - Most Expensive



Garage Type (1-6)

- Least - Most Value



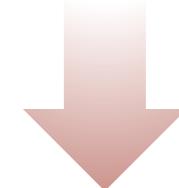
Residential Density (1-4)

- Lowest - Highest Price per Sq. Ft.

Ranking Quality/Condition

- ▶ Exterior Quality
- ▶ Exterior Condition
- ▶ Heating Quality
- ▶ Kitchen Quality
- ▶ Fireplace Quality
- ▶ Garage Quality
- ▶ Garage Condition

Excellent
Good
Avg./Typical
Fair
Poor



5
4
3
2
1

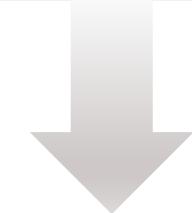
One-Hot Encoding

- ▶ Building Type
 - Single-family Detached
 - Townhouse
- ▶ Year Sold
- ▶ Zone Type
 - Commercial
 - Industrial
- ▶ Proximity Features
 - Near Amenity
 - Near Road/Rail



Calculated Fields

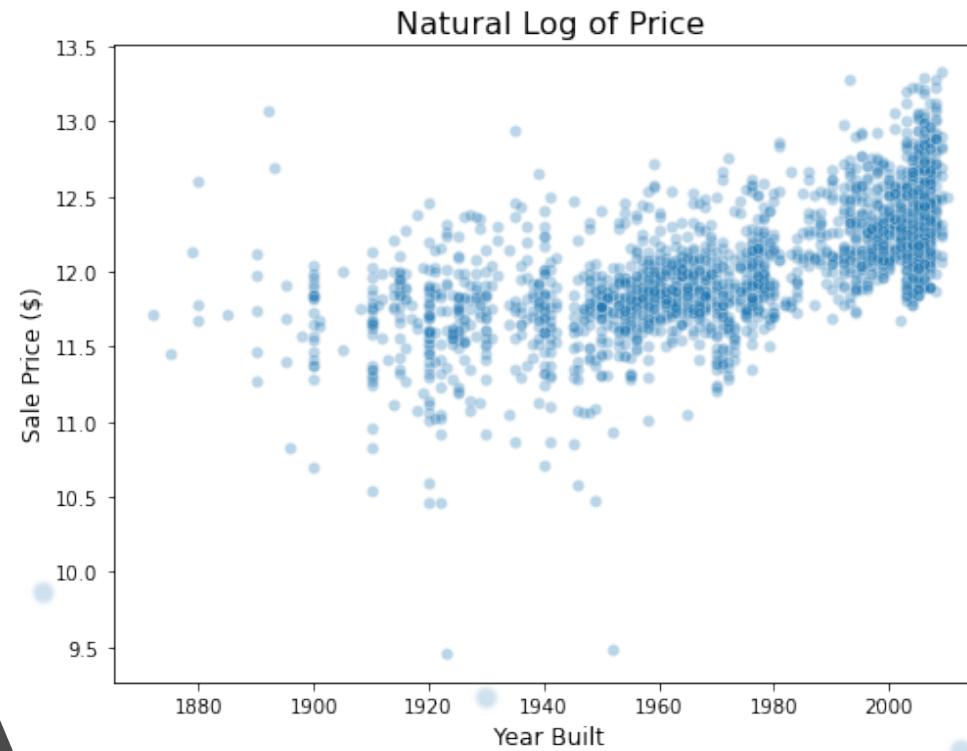
Total Basement Sq Ft. + 1st Floor Sq Ft. +
2nd Floor Sq Ft.



Total Square
Footage

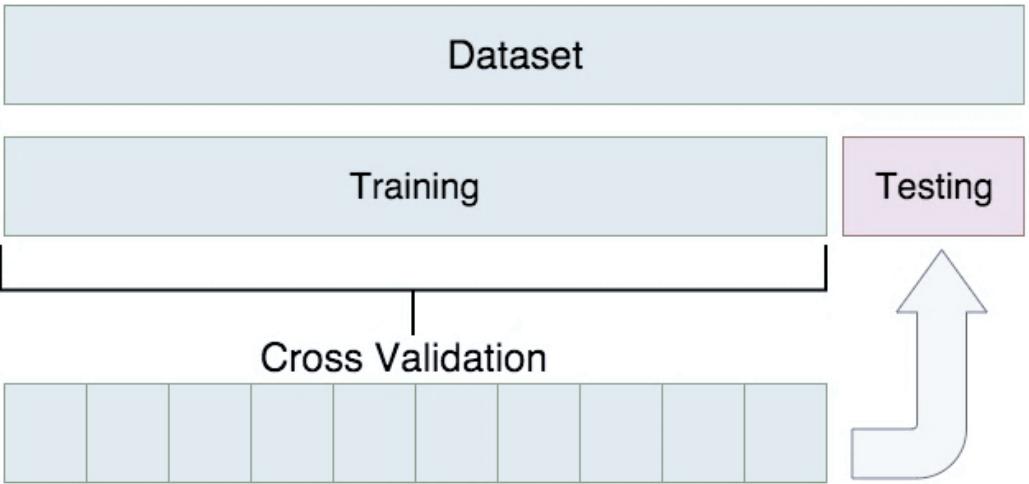
Natural Log of Price

- ▶ Reduces the range of values
- ▶ Brings distribution to a more linear shape



Part III : The Model

Fit, transform, and score several Regression models



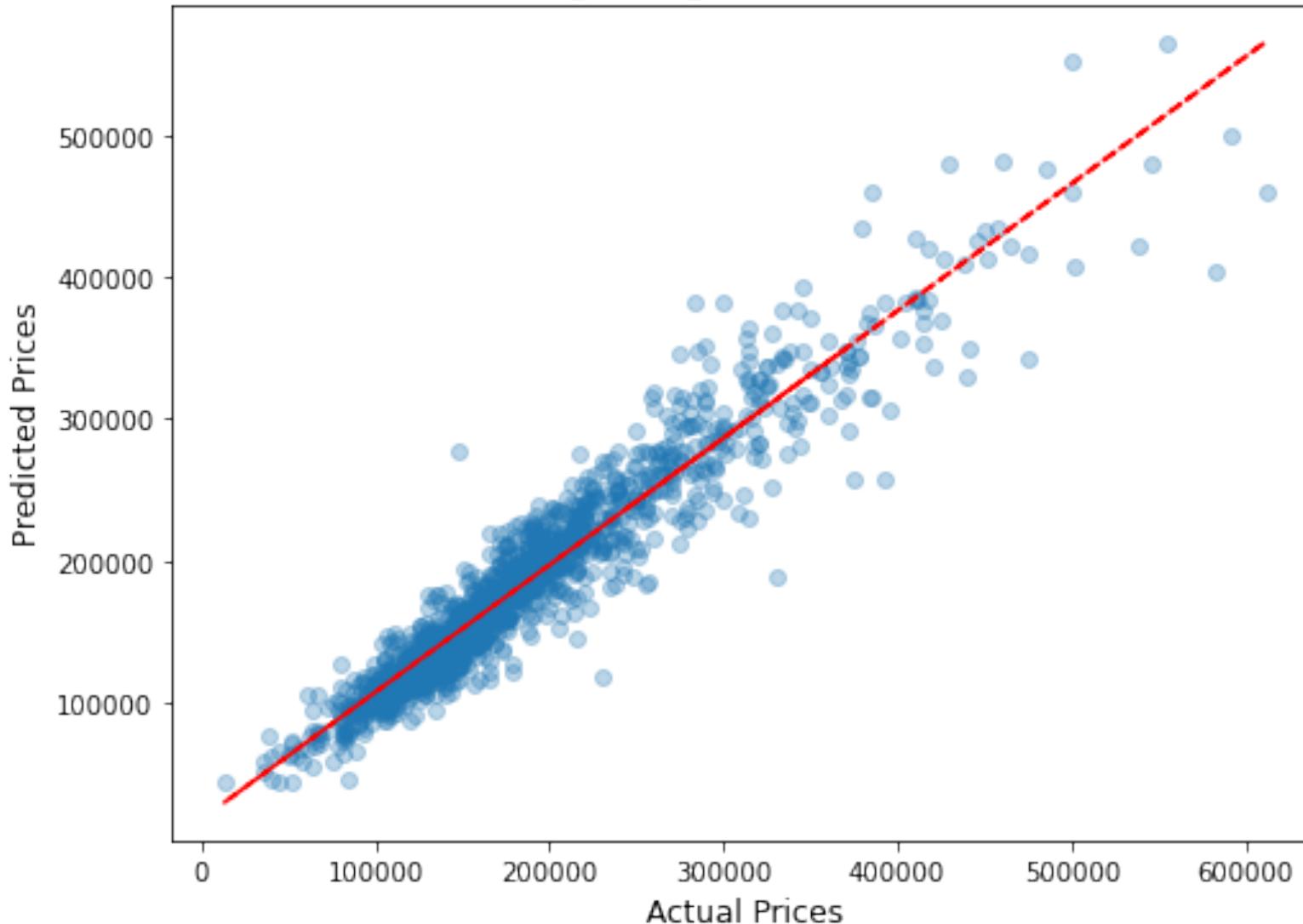
Segment Data

- ▶ Training Group (75%)
 - 1538 observations
- ▶ Testing Group (25%)
 - 513 observations
- ▶ Unknown
 - 878 observations where price is not given

The baseline model (mean)
has an average error of \$78,194

	Linear Regression	Ridge Regression ($\alpha = 18.48$)	Lasso Regression ($\alpha = .01$)
Training Set ▼			
R ² Score	0.9084	0.9084	0.9084
Cross Validation	0.9017	0.9018	0.9017
Average Error	\$23,594	\$23,629	\$23,587
Testing Set ▼			
R ² Score	0.8244	0.826	0.8249
Average Error	\$36,526	\$35,830	\$36,522

Ridge Regression Model



Feature	Unit Change
Total Sq Ft.	\$27,775
Overall Quality	\$8,516
Year Built	\$6,867
Fireplaces	\$6,102
Neighborhood	\$5,546
Garage Cars	\$5,231
Kitchen Qual	\$3,989
Heating QC	\$2,660
Residential	\$2,583
Wood Deck	\$1,560
Single Family	\$1,479
Garage Type	\$1,418
Near Amenity	\$1,187
Full Baths	\$1,180
Bedrooms	\$787
Industrial	\$330
Duplex	-\$672
Near Rail	-\$770
Townhouse	-\$1,068
Near Road	-\$1,385
Commercial	-\$3,916

Areas for Further Research

- ▶ More detail on Kitchen finishes (countertops, backsplashes) and appliances
- ▶ More information on storage square footage (closets)
- ▶ Include information about local school districts
- ▶ Clarify “positive off-site features”