Steven Wang
Final Project: Feb. 11, 2017
Ames Housing Dataset

Problem: Predict the sale price of homes, using the features of each property.

Hypothesis: The sale price of a home is positively correlated with (1) its size and (2) how recently it was built.

Ames Housing Dataset

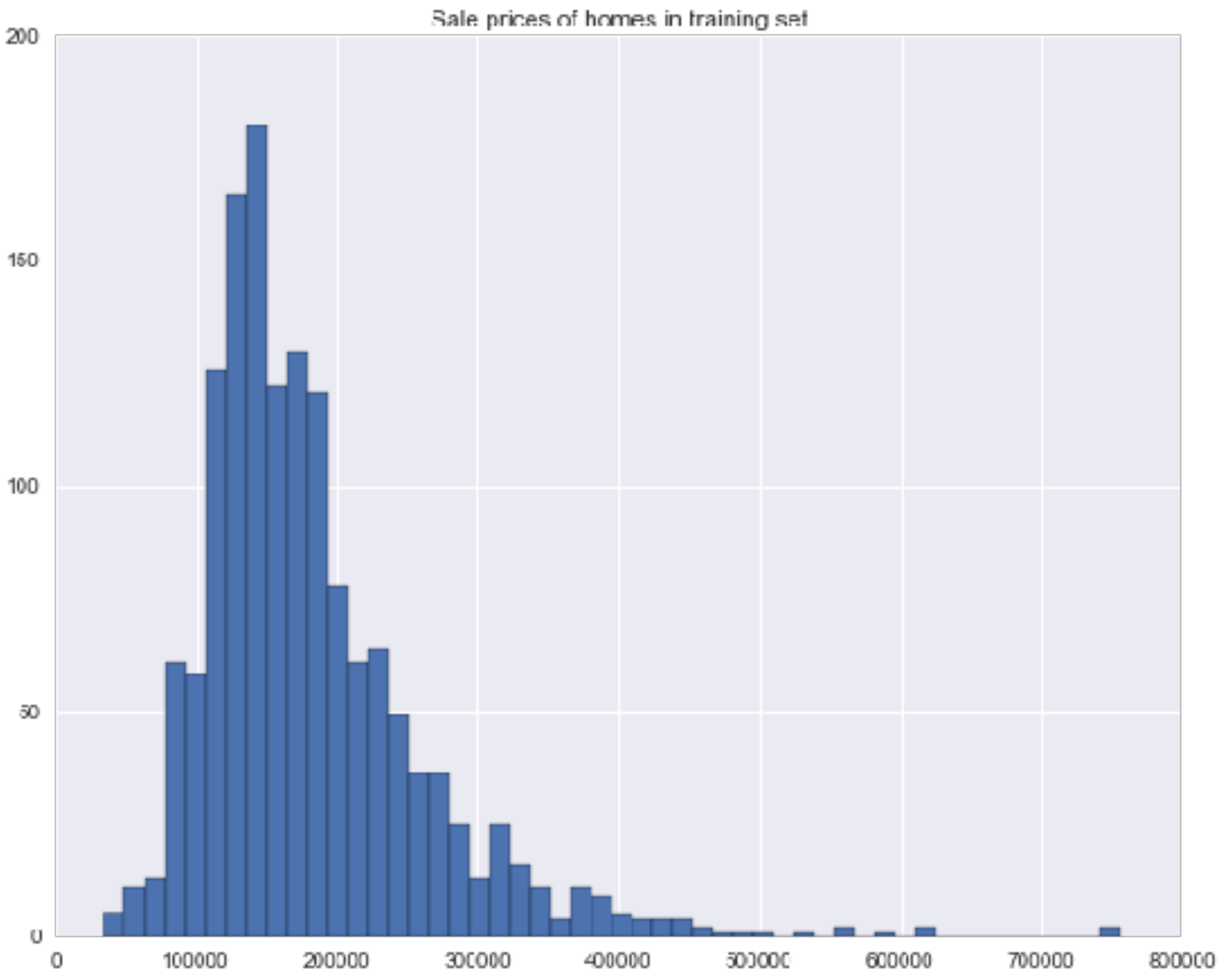
- 79 features of 2920 residential homes sold in Ames, Iowa, from 2006 through 2010.
  - Type of home (one-story, two-story, etc.)
  - Dimensions of home (lot size, living area, garage size, basement size, etc.)
  - Condition of home
  - Year home was built
  - Neighborhood
  - Other features (fireplace? pool? etc.)
- Kaggle splits the dataset into training and test sets of equal size, i.e., 1460 observations each.

Ames, Iowa

2010 Census	
White	85%
Asian	9%
African American	3%
Hispanic/Latino	3%

- Ranked 9th on CNN's "Best Places to Live" in 2010
- Total population: 58,965, as of 2010 Census
- Median family income: $56,439
- Largest employer: Iowa State University

Exploratory Data Analysis: Target (Sale Price)



Mean: $180,921

Std Dev: $79,442

Min: $34,900

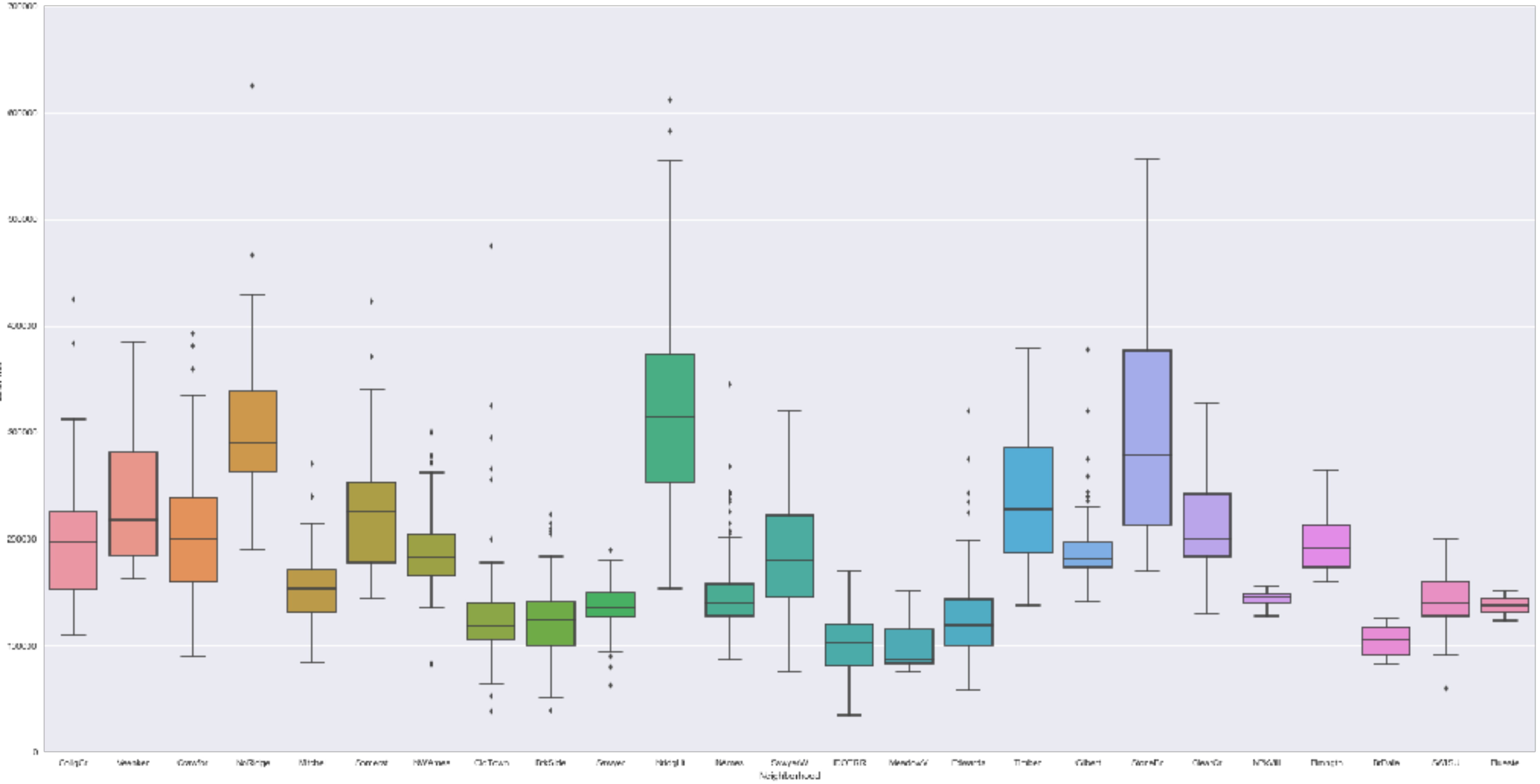
25%: $129,975

50%: $163,000

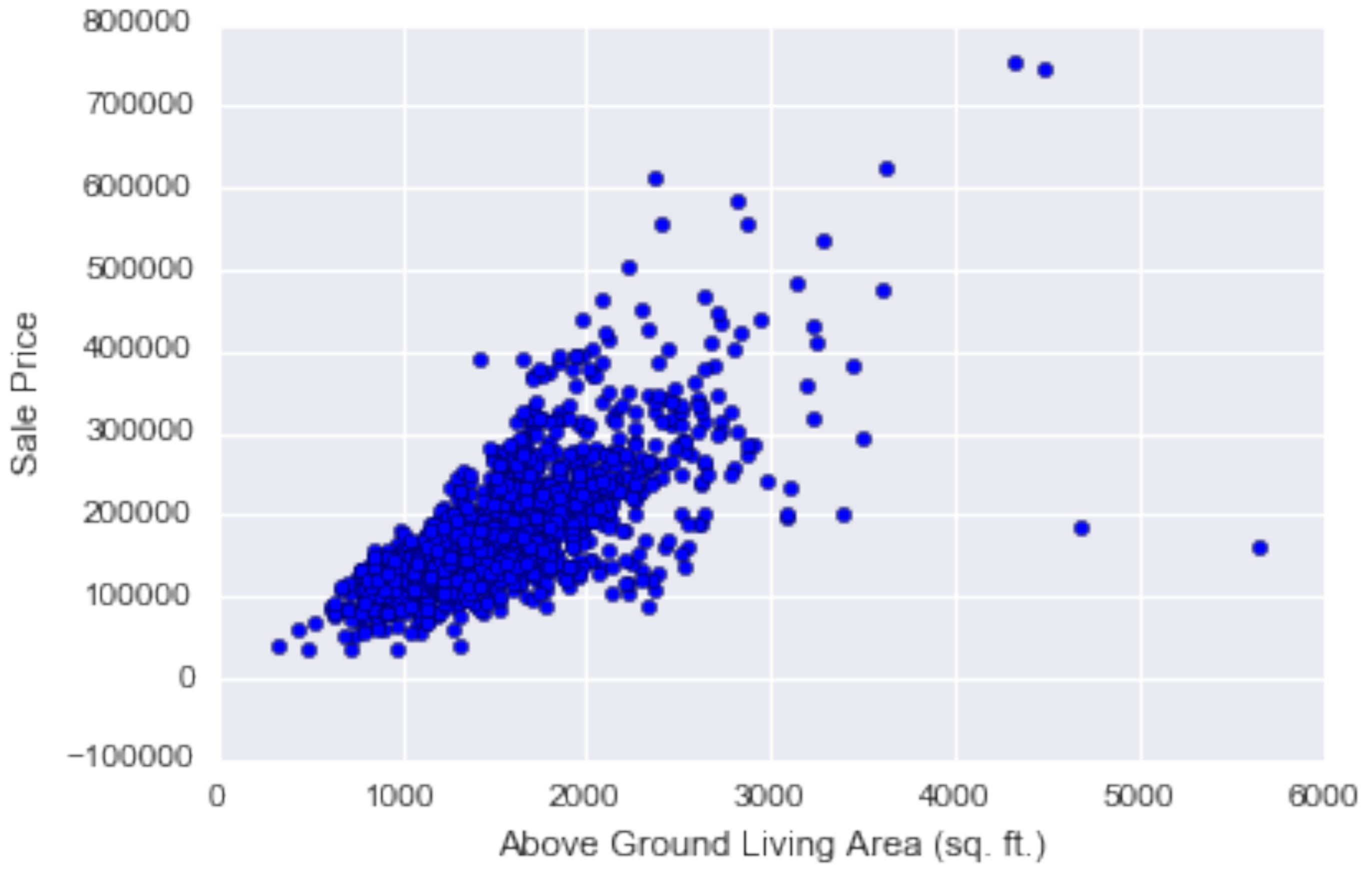
75%: $214,000

Max: $755,000

Sale Prices by Neighborhood



Exploratory Data Analysis: Above-Ground Living Area



Outliers: Homes with more than 4000 square feet of above-ground living area were dropped from the training set.

Mean: 1515 sq. ft.

Std Dev: 525 sq. ft.

Min: 334 sq. ft.

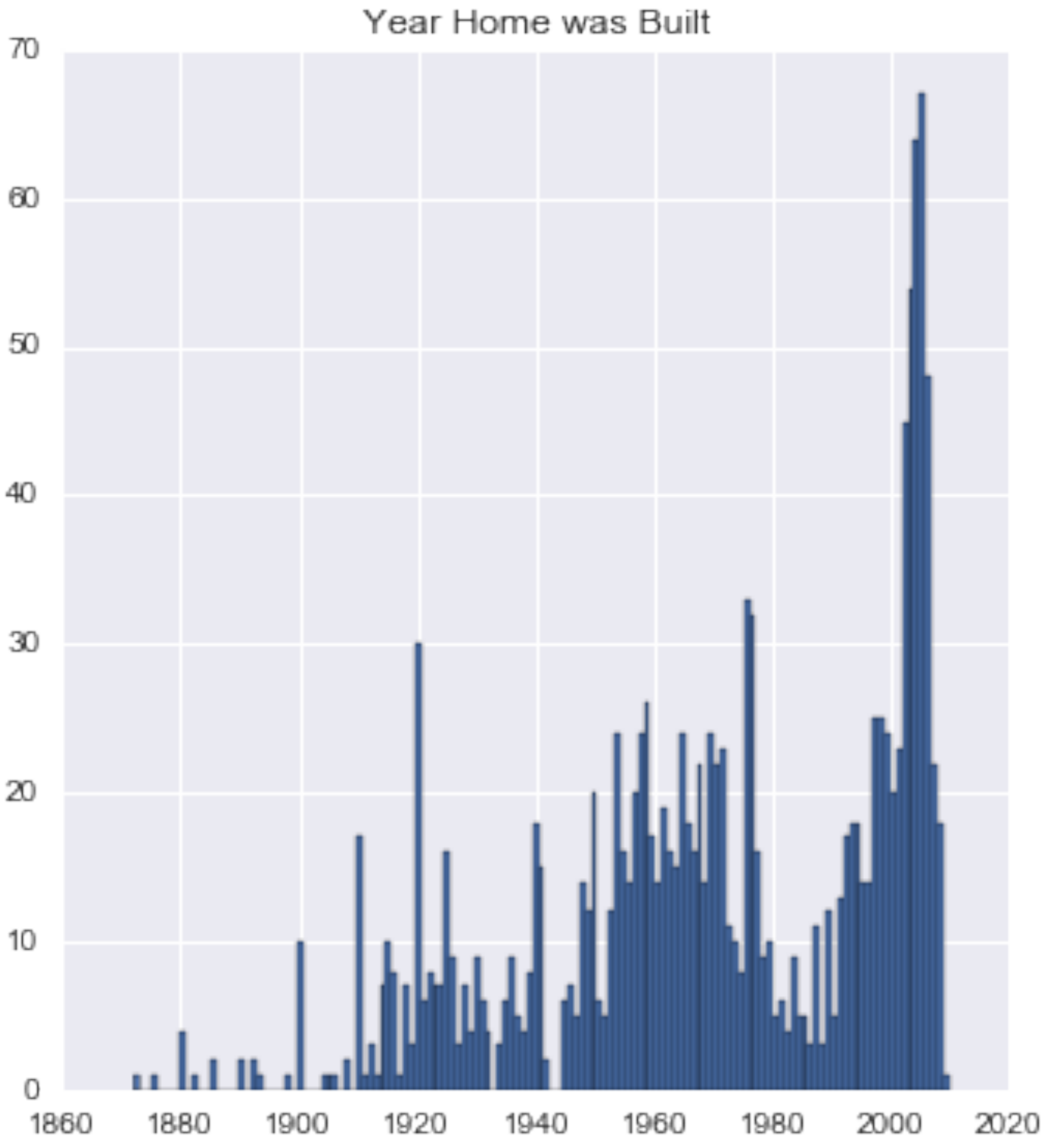
25%:1129.5 sq. ft.

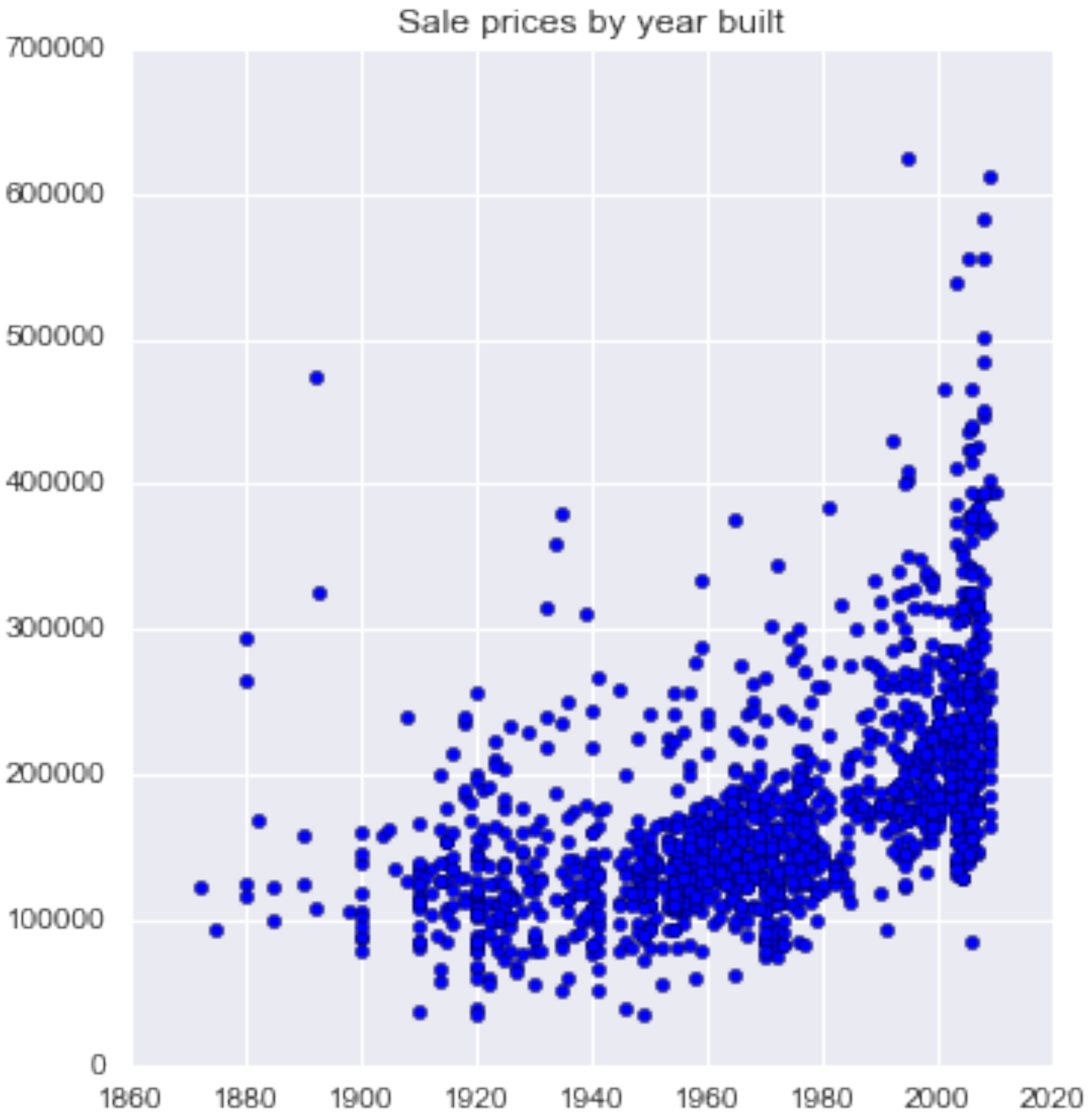
50%: 1464 sq. ft.

75%: 1776.75 sq. ft.

Max: 5642 sq. ft.

Exploratory Data Analysis: Year of Construction

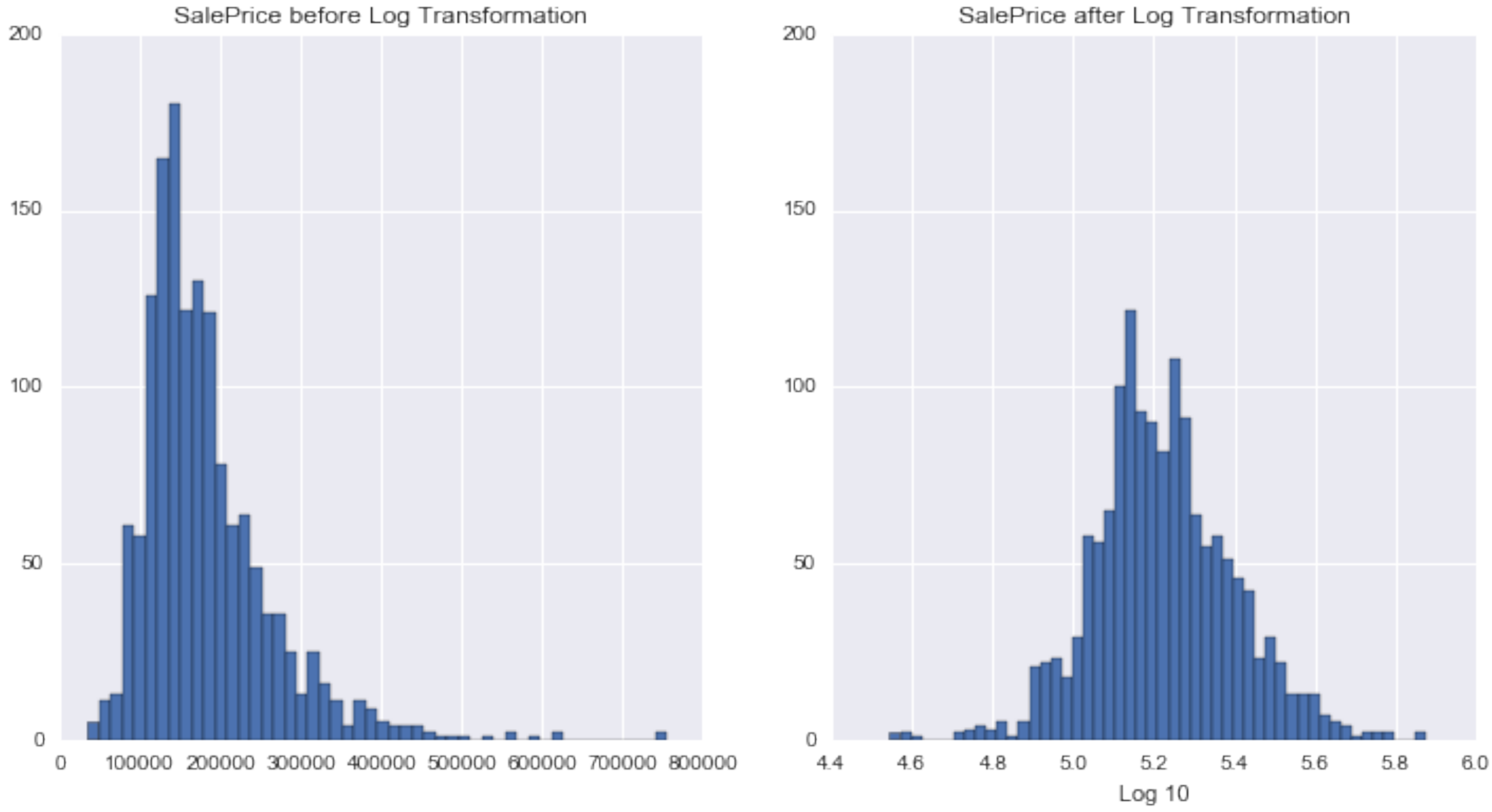




Feature Engineering

- Missing Values
- Dummy Variables: Resulted in 296 features
- Skewed Features: Logarithmic and Box-Cox transformation

Sale Price, before and after Logarithmic Transformation



Baseline

- Root Mean Squared Error (RMSE)
- If mean sale price ($181K) is predicted for every row in training set, RMSE = $76,670

Linear Regression, Lasso, Ridge, and Random Forest

Model	RMSE	Notes
Linear Regression (“Kitchen Sink”)	$78,919	Worse than baseline!
Lasso	$18,518	Reduced features from 296 to 118
Linear Regression (Lasso Features)	$17,599	
Ridge	$16,365	
Random Forest	$10,571	Top 11 features explain 85%

Random Forest Feature Importance

	Feature	Importance
1	Above-Ground Living Area	0.3419
2	Year Built	0.1645
3	Quality of Exterior Materials (Typical)	0.0924
4	Total Basement Area	0.0780
5	Total Garage Area	0.0712
6	Central A/C?	0.0245
7	Basement Finished Area (sq. ft.)	0.0223
8	Lot Area	0.0206
9	Year Garage Was Built	0.0134
10	Home Does Not Have Fireplace	0.0115
11	Remodeling Date	0.0114
	Total	0.8518

Kaggle Scores

1,722 out of 4,206 (as of 2/10/2017)

Model	**RMSLE**
Lasso	0.12832
Linear Regression (with Lasso Features)	0.12672
Ridge	0.12980
Random Forest	0.15151

Next Steps

- Location, location, location?
- Better Kaggle results?
  - Top 50: RMSLE of ~ 0.11
  - More feature engineering?
  - XGBoost? Other models?