Ames Housing

2006-2010 Review

How did Ames get its name?

The city was named after 19th century U.S. congressman Oakes Ames of Massachusetts.

City of Ames

- 9th on CNN Money's 2010 list of Best Places to Live in the US
- Median home price in 2010: $175,000
- Population of about 60,000
- Median age of 26
- Home to Iowa State University

How do we determine sale price from a list of features?

Ames Housing Data - Metrics

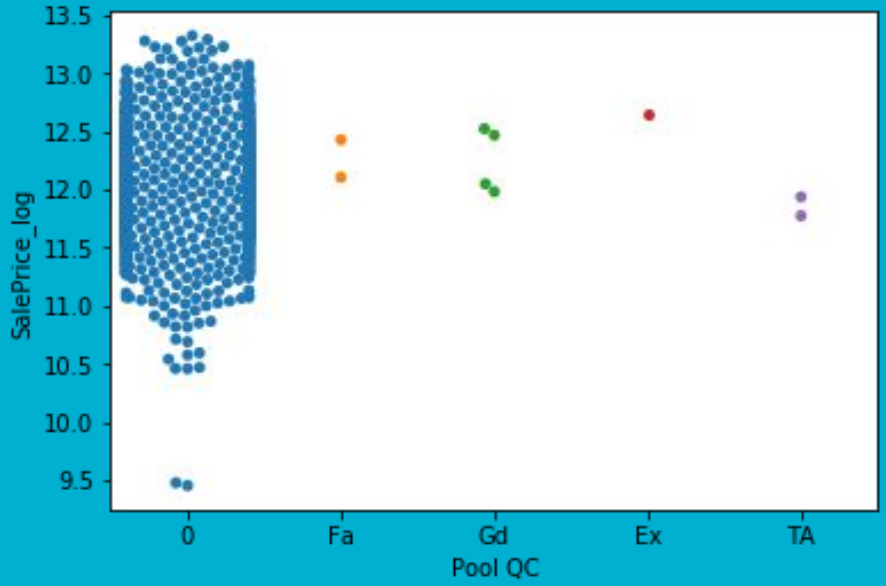
Averages of 2,051 Listings

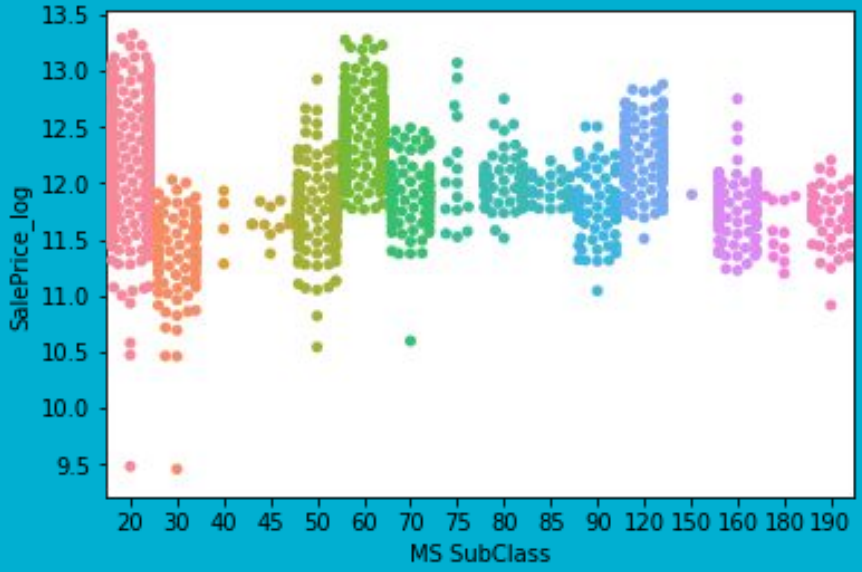
- Lot Area - 10,000 Square Feet
- Year Built - 1972
- Above Ground Living Area - 1,500 Square Feet
- 2 Baths
- 3 Bedrooms
- 2 Car Garage
- Pool Area - 2 Square Feet
- Sale Price - $181,470

List of Features

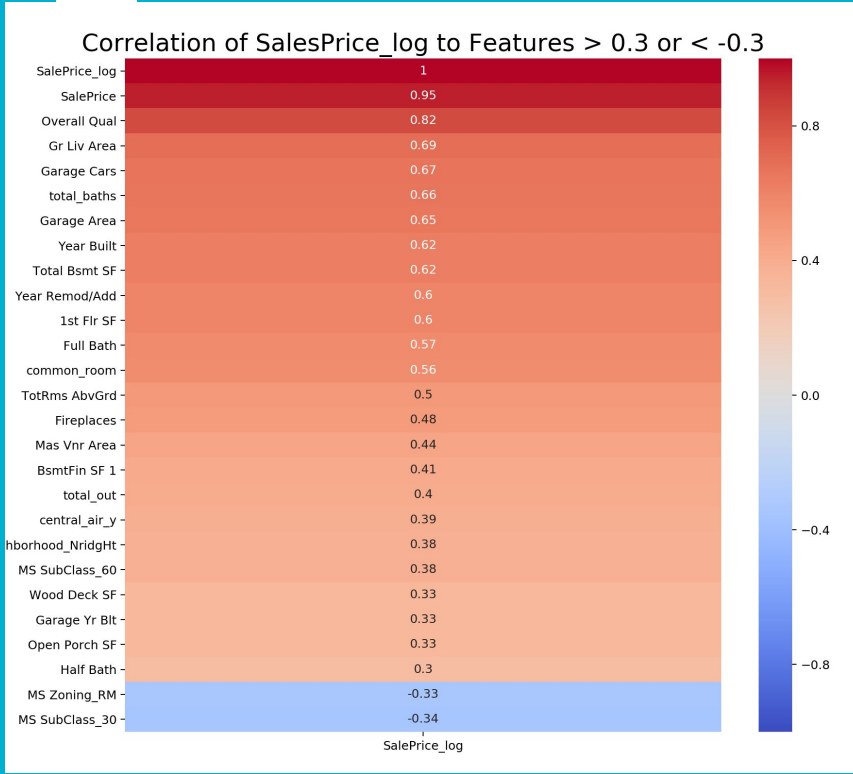
Id	Bldg Type	Bsmt Cond	Bsmt Full Bath	Garage Qual
PID	House Style	Bsmt Exposure	Bsmt Half Bath	Garage Cond
MS SubClass	Overall Qual	BsmtFin Type 1	Full Bath	Paved Drive
MS Zoning	Overall Cond	BsmtFin SF 1	Half Bath	Wood Deck SF
Lot Frontage	Year Built	BsmtFin Type 2	Bedroom AbvGr	Open Porch SF
Lot Area	Year Remod/Add	BsmtFin SF 2	Kitchen AbvGr	Enclosed Porch
Street	Roof Style	Bsmt Unf SF	Kitchen Qual	3Ssn Porch
Alley	Roof Matl	Total Bsmt SF	TotRms AbvGrd	Screen Porch
Lot Shape	Exterior 1st	Heating	Functional	Pool Area
Land Contour	Exterior 2nd	Heating QC	Fireplaces	Pool QC
Utilities	Mas Vnr Type	Central Air	Fireplace Qu	Fence
Lot Config	Mas Vnr Area	Electrical	Garage Type	Misc Feature
Land Slope	Exter Qual	1st Flr SF	Garage Yr Blt	Misc Val
Neighborhood	Exter Cond	2nd Flr SF	Garage Finish	Mo Sold
Condition 1	Foundation	Low Qual Fin SF	Garage Cars	Yr Sold
Condition 2	Bsmt Qual	Gr Liv Area	Garage Area	SalePrice

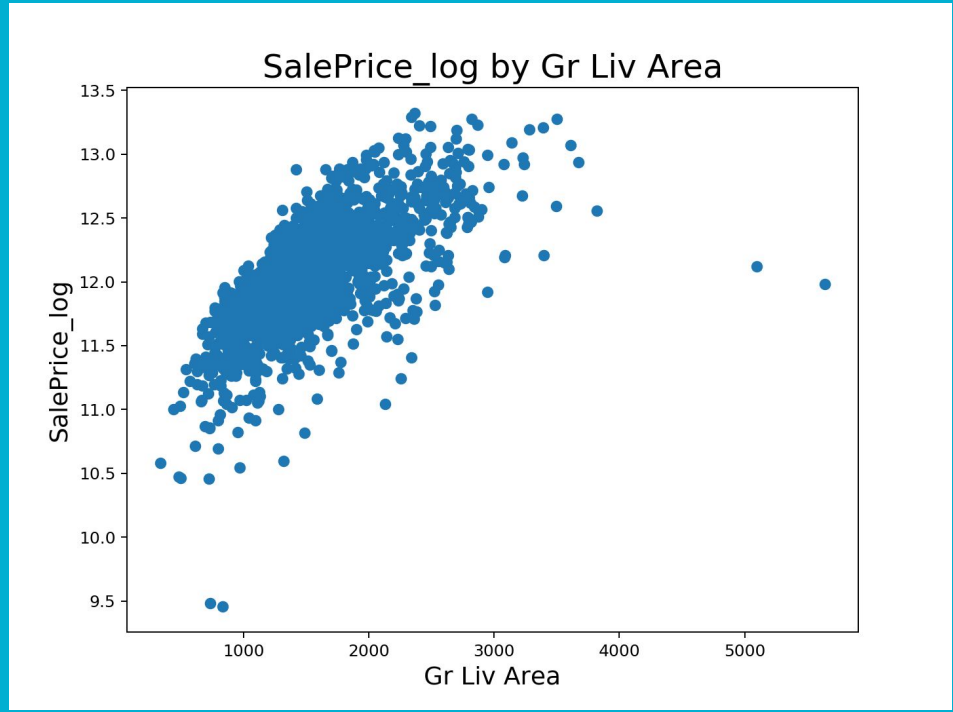
Swarm Plots for Categorical Features





Correlation and Scatter Plots for Numeric



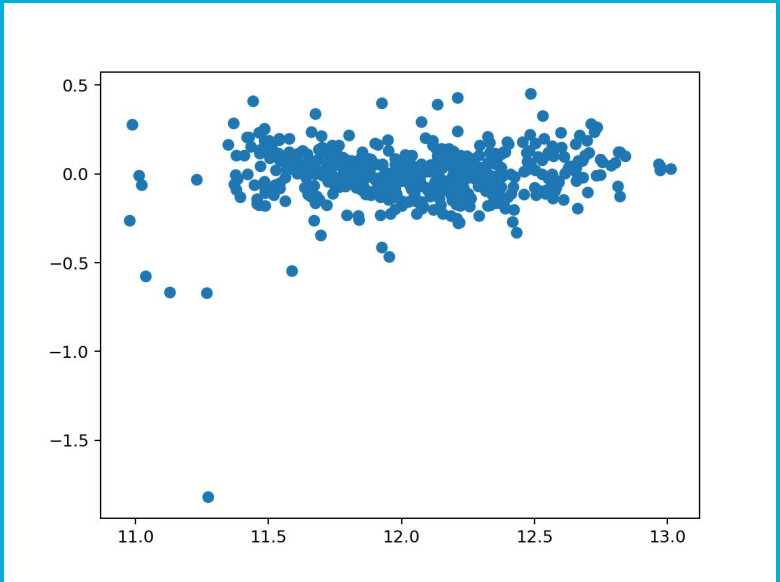


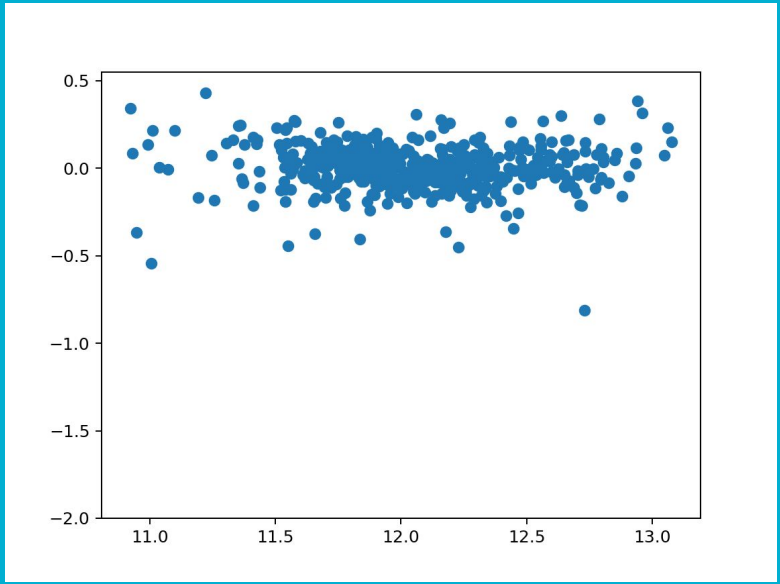
Starting Features Selected

- Overall Qual
- Gr Liv Area
- Garage Cars
- Total_baths
- Year Remod/Add
- Total Bsmt SF
- Common_room
- Fireplaces
- Total_out
- Lot Area
- MS Zoning
- Neighborhoods

I selected 10 numeric features that had high correlations with Sales Price, and 2 categorical features as a start.

Residuals





The plots are evenly distributed, with a few outliers. Doesn't appear to be a pattern, so homoscedastic. The graph on the left is from an earlier attempt, and the graph on the right is from my final attempt. The distribution isn't as broad on the right, so I've reduced my error.

Model Iterations

• Added new log transformed features that look normally distributed: 'Gr Liv Area_log', 'total_baths_log', and 'Lot Area_log'	Higher train scores, but lower Kaggle score
• Added month and year sold	Lower train and Kaggle score
• Fit the model on the entire set of training data	Higher Kaggle score
• Used PolynomialFeatures to fit the model.	Higher train, but lower Kaggle
• Added 'street_paved', 'central_air_y' as columns. Dummied 'Land Contour', 'MS SubClass', 'Functional', 'Exter Qual', 'Condition 1' and 'Condition 2' columns. Added 'Bedroom AbvGr', 'Kitchen AbvGr', and 'Misc Val' as features.	Higher train and Kaggle score

Sales Price Determination Factors

- There are a lot of features that go into determining the price!
- Steps to select the best set of features:
  - Evaluate how much data is within each feature
  - Use boxplots or swarmplots to see how price clusters around categorical features
  - Refer to correlation to narrow down the list of numeric features
  - Don’t include features that overlap
  - Think like a buyer!