The Problem

How do we predict the price of a home in Ames, Iowa?

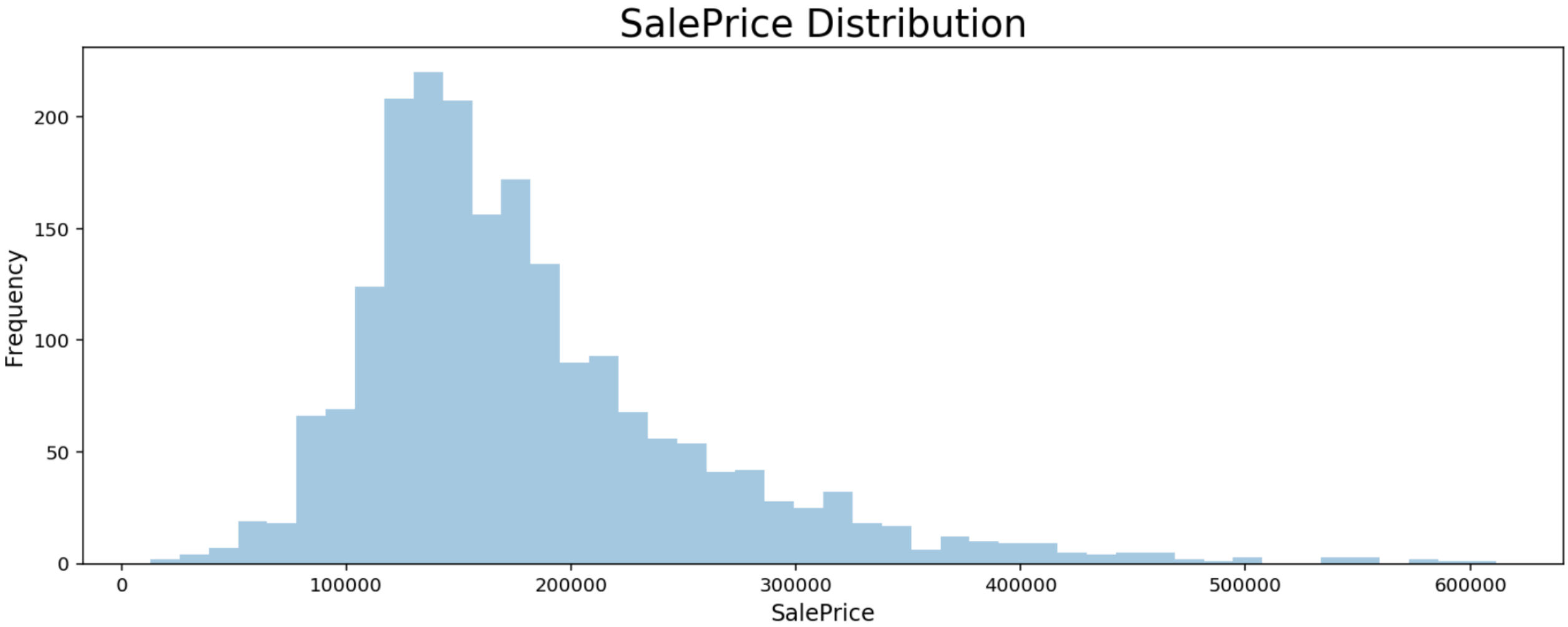
The Issue in Context

- Don't want to pay rent every month for the rest of your life? Buy your own home!
- Tired of living in a small town? Sell your home!
- But…how do you put a price tag on it?

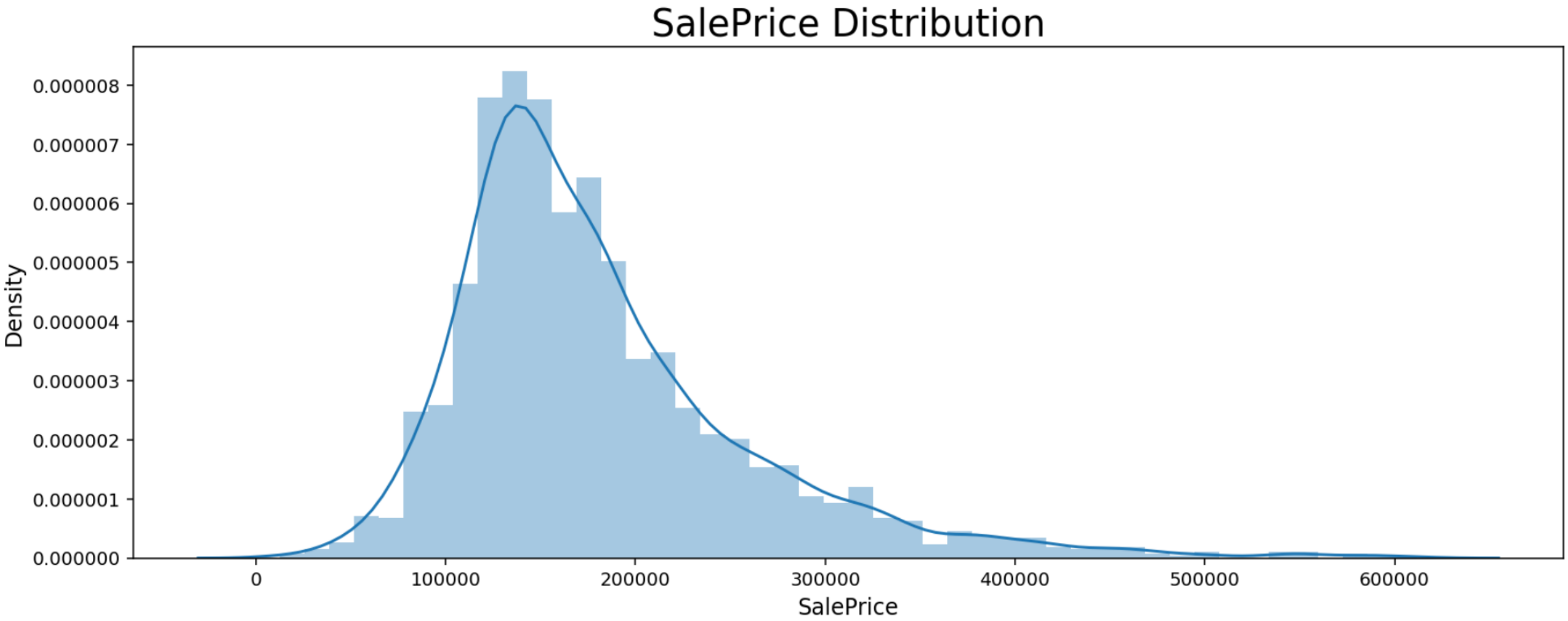
The Approach

- Explore large set of data on over 2000 homes in Ames, IA
- Construct *predictive* linear regression model
- Evaluate model on unseen data

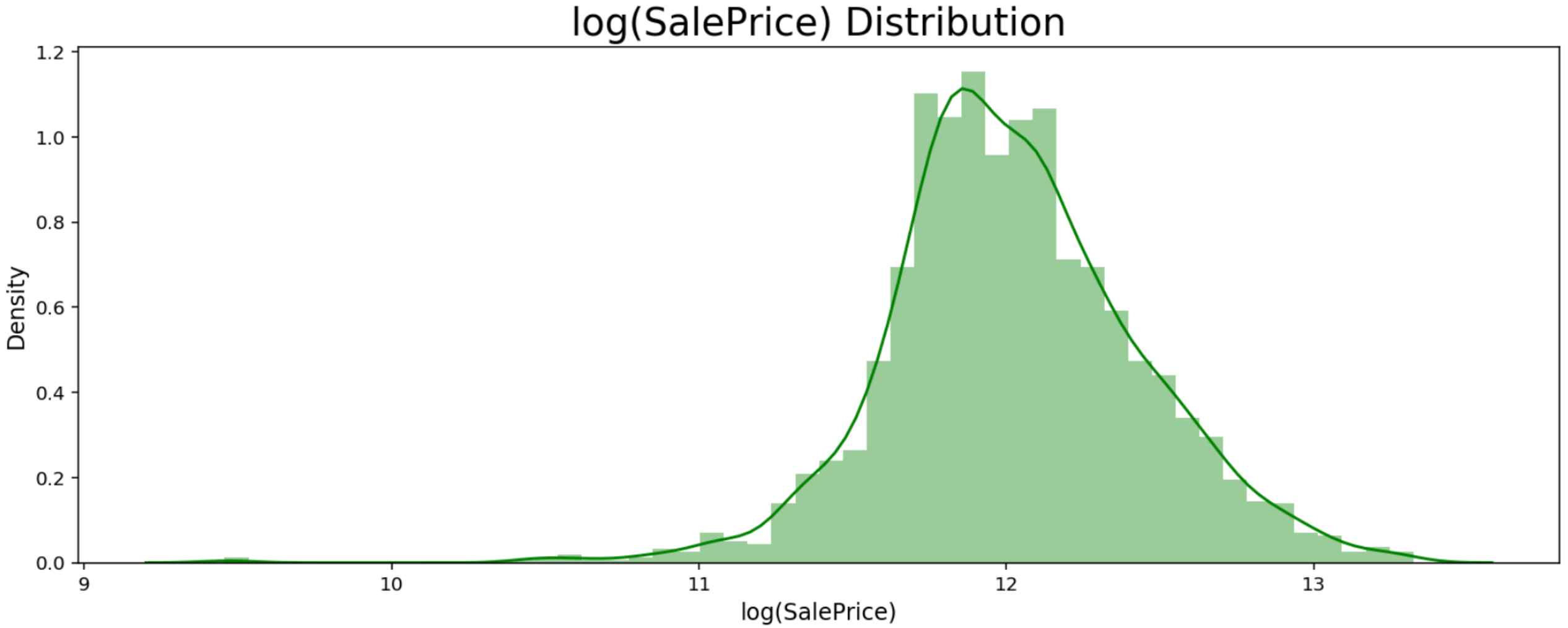
Examining the Distribution of Target Variable – *SalePrice*



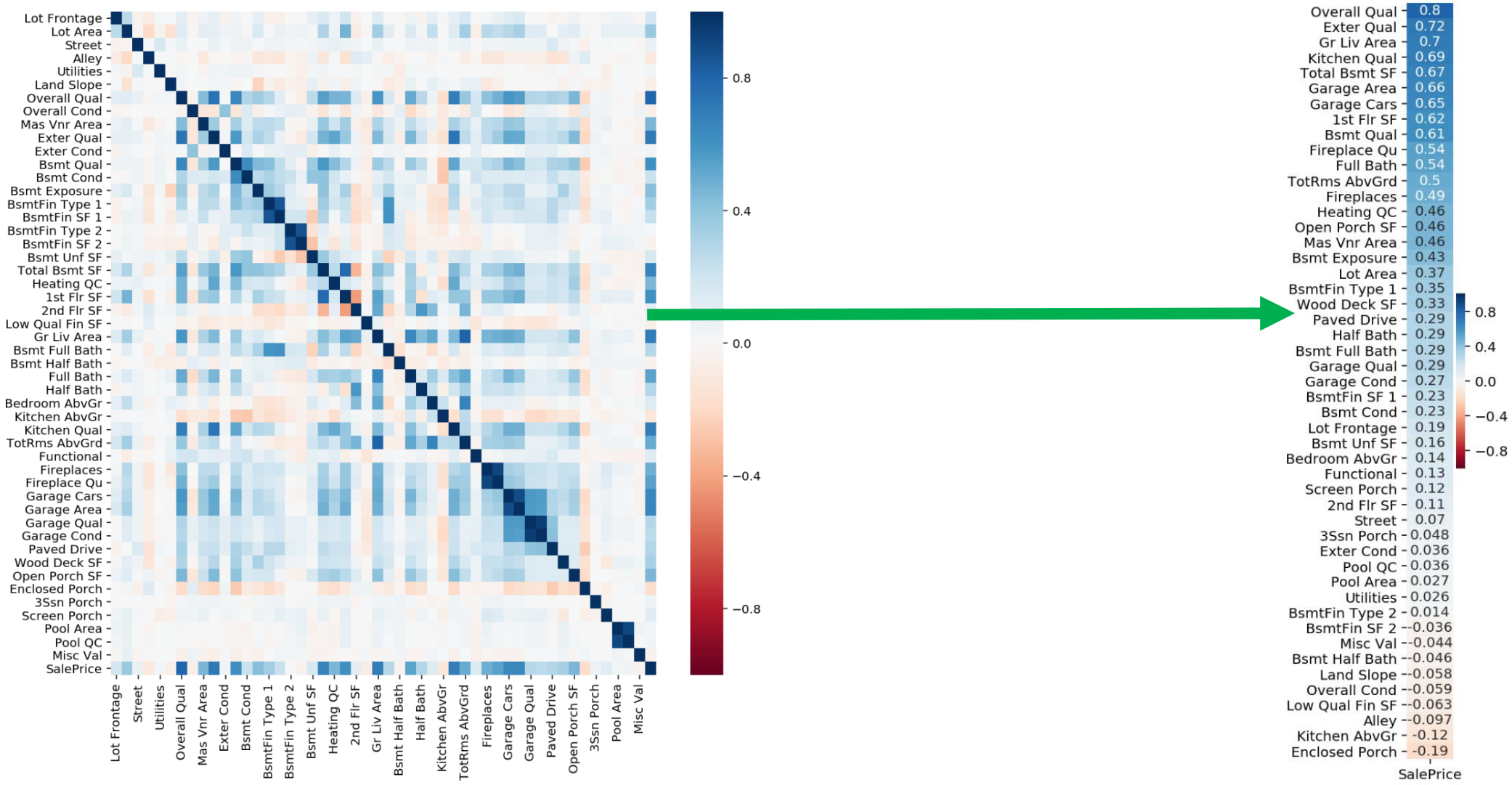
Examining the Distribution of Target Variable – *SalePrice*



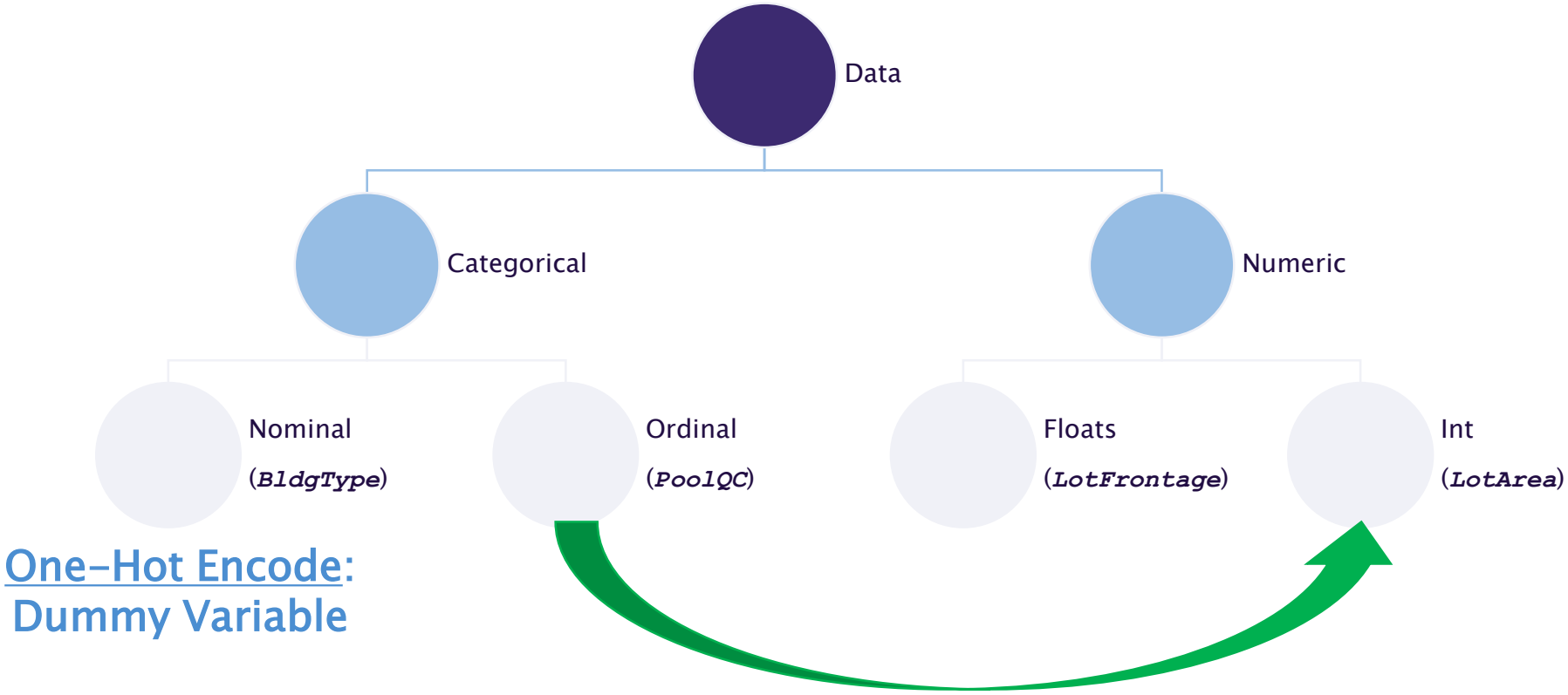
Examining the Distribution of Target Variable – log(*SalePrice*)



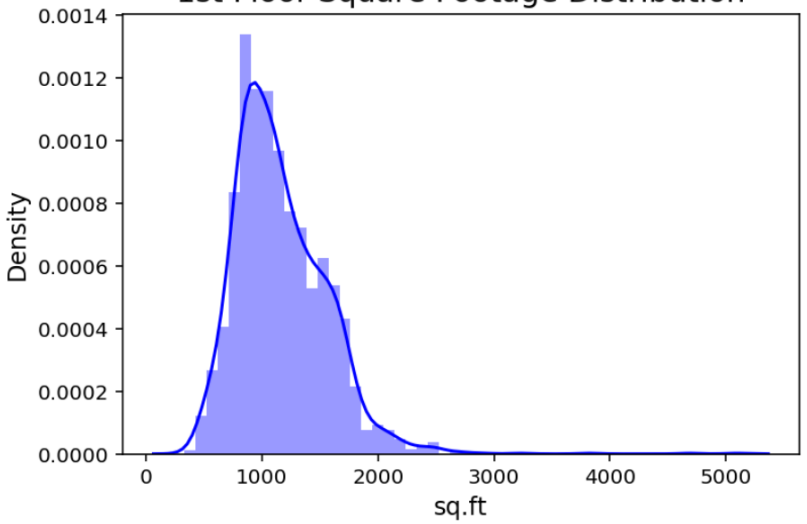
Some multicollinearity in features, but focus on correl. vs *SalePrice*

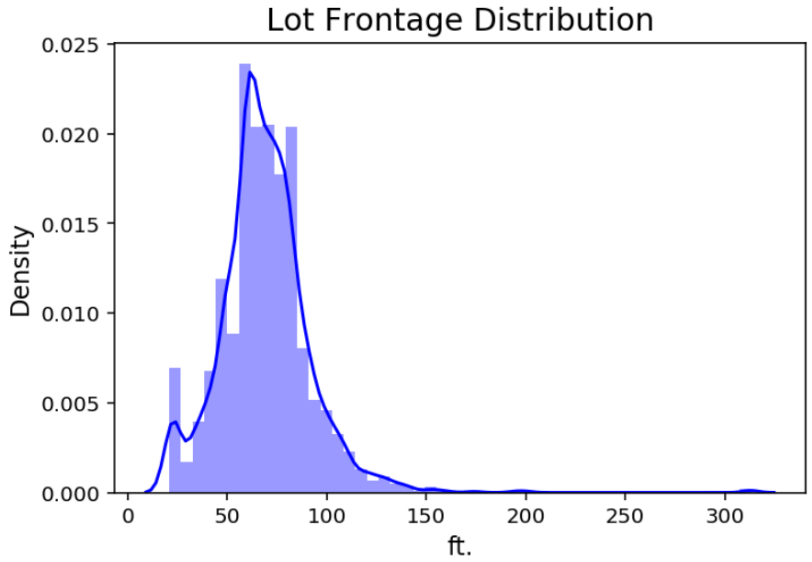


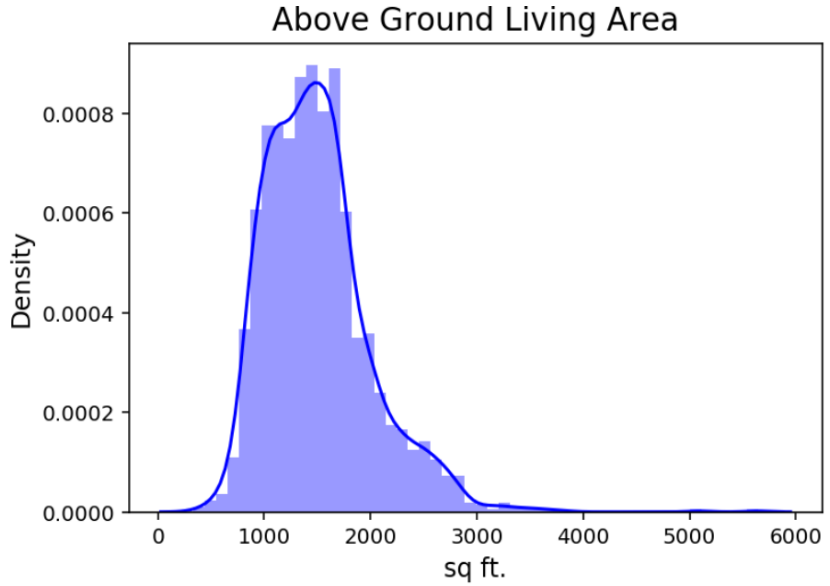
Handling Features by Data Type



Distributions Some of Noteworthy Numeric Features







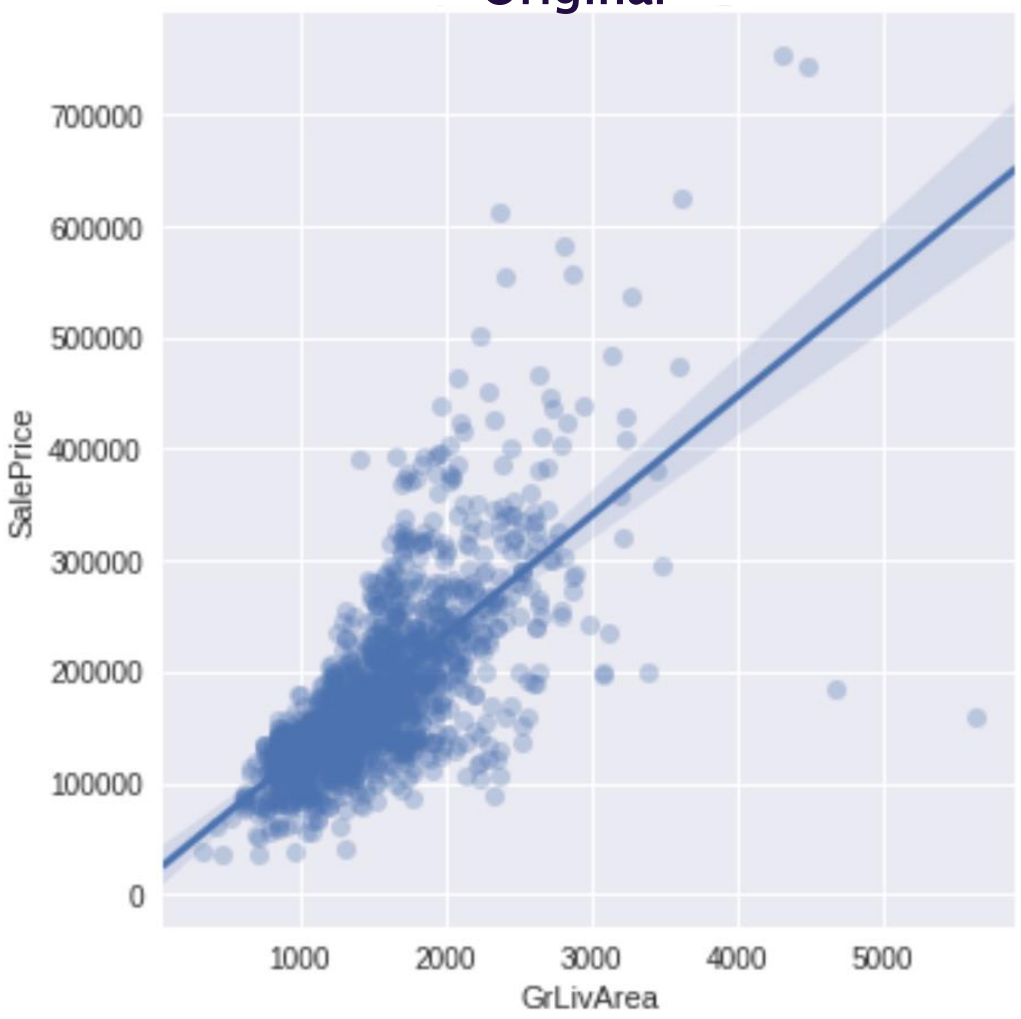
There are also features that are "almost normal" but skewed in the same way that ***`SalePrice`*** is!

Our threshold: a feature is moderately skewed if Skew(X) > 0.5

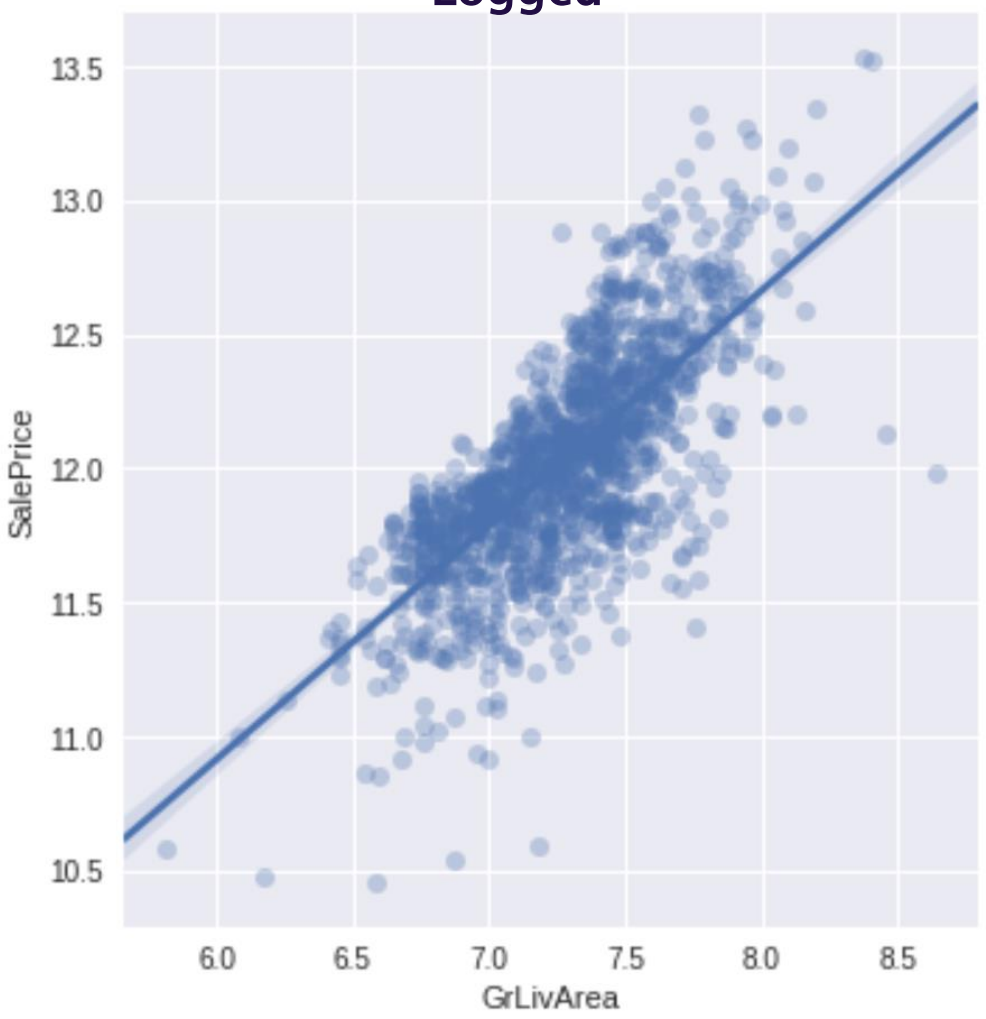
Not convinced? See next slide for evidence to support replacing such features with their logs.

Benefits of the Log Transform

Original

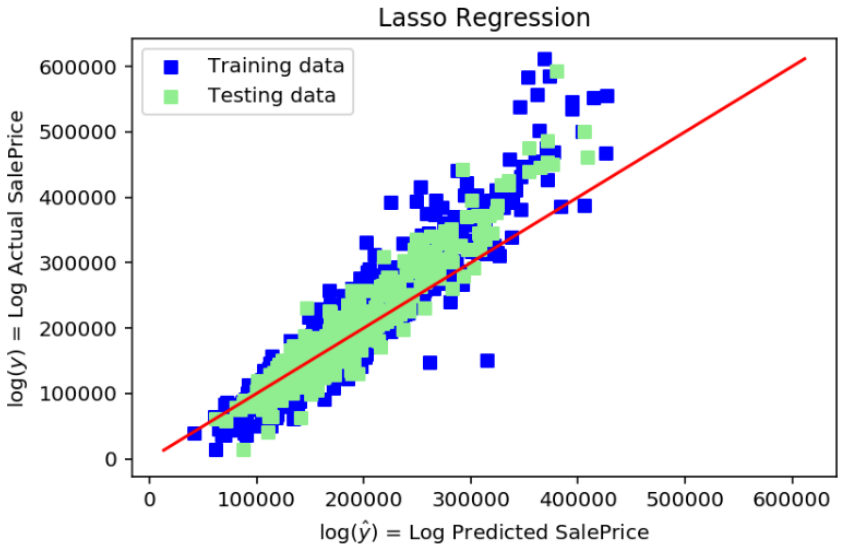


Logged



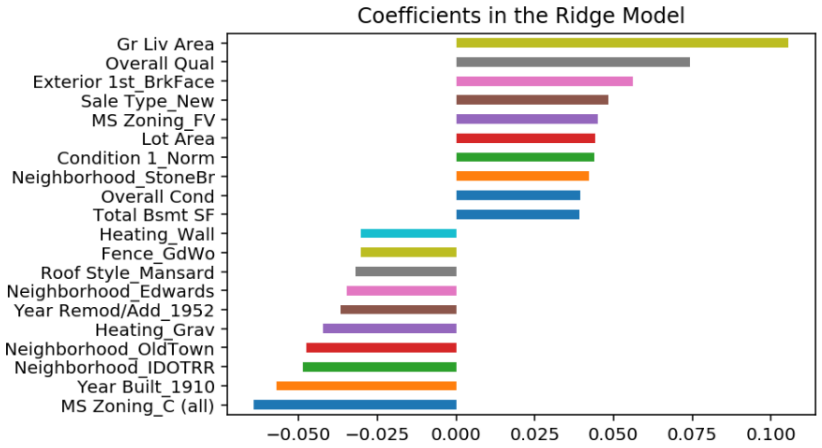
Modeling: Results

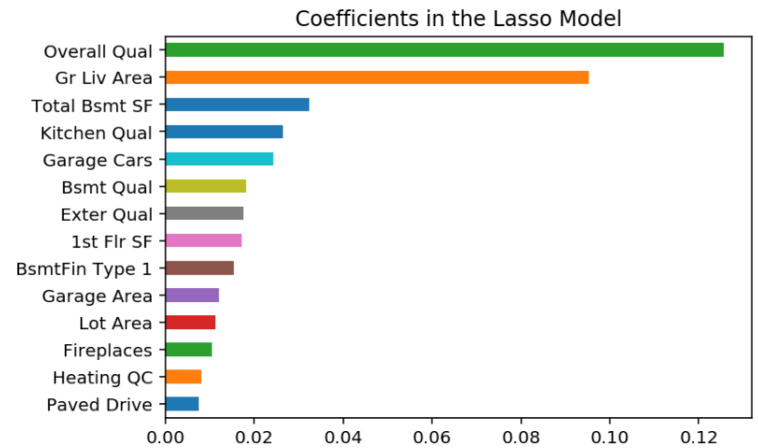
OLS useless with 516 regressors, but Ridge and Lasso came through



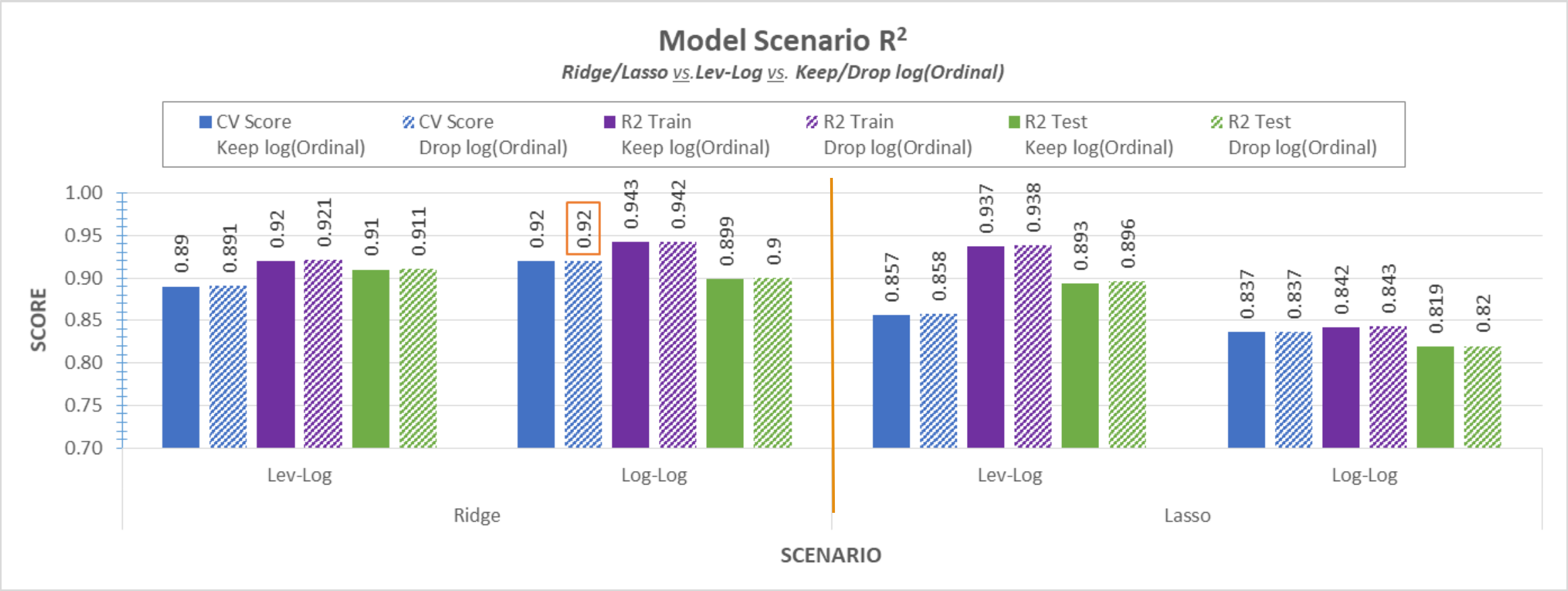
Ridge picked 490 features and ignored the other 26 features.

Lasso picked 14 features and zero-ed out the other 502 features.





Scenario Analysis of Models



- **Log-Log Ridge Regression** outperformed other scenarios in terms of both CV Score and Training Set R^2.
- **Dropping log(ordinals)** typically did better than keeping them in all scenarios.

Verdict on our Log-Log Ridge Regression model

- Able to handle unseen data decently well
  - ➢ R^2 Test score, Kaggle RMSE score
- Suffers from slight overfitting *(5% diff in R^2 Train vs. Test),*
- Still performs well on a *consistent* basis judging by 5-Fold Cross Validation Score
- Overall, model can be used to predict prices of homes in Ames, IA within a ballpark of true value.

Further Enquiry

- Manually select interaction terms to include in the model and examine if RMSE falls further
- Introduce polynomial versions of the top 10 highly correlated features with SalePrice into the model to see if it improves model predictability
- Implement Gradient Boosting and Random Forests to improve model predictability
- Using the surviving features from Lasso, how well would OLS perform? What ends up being the threshold at which OLS fails?