Predicting Home Sale Price in Ames, IA

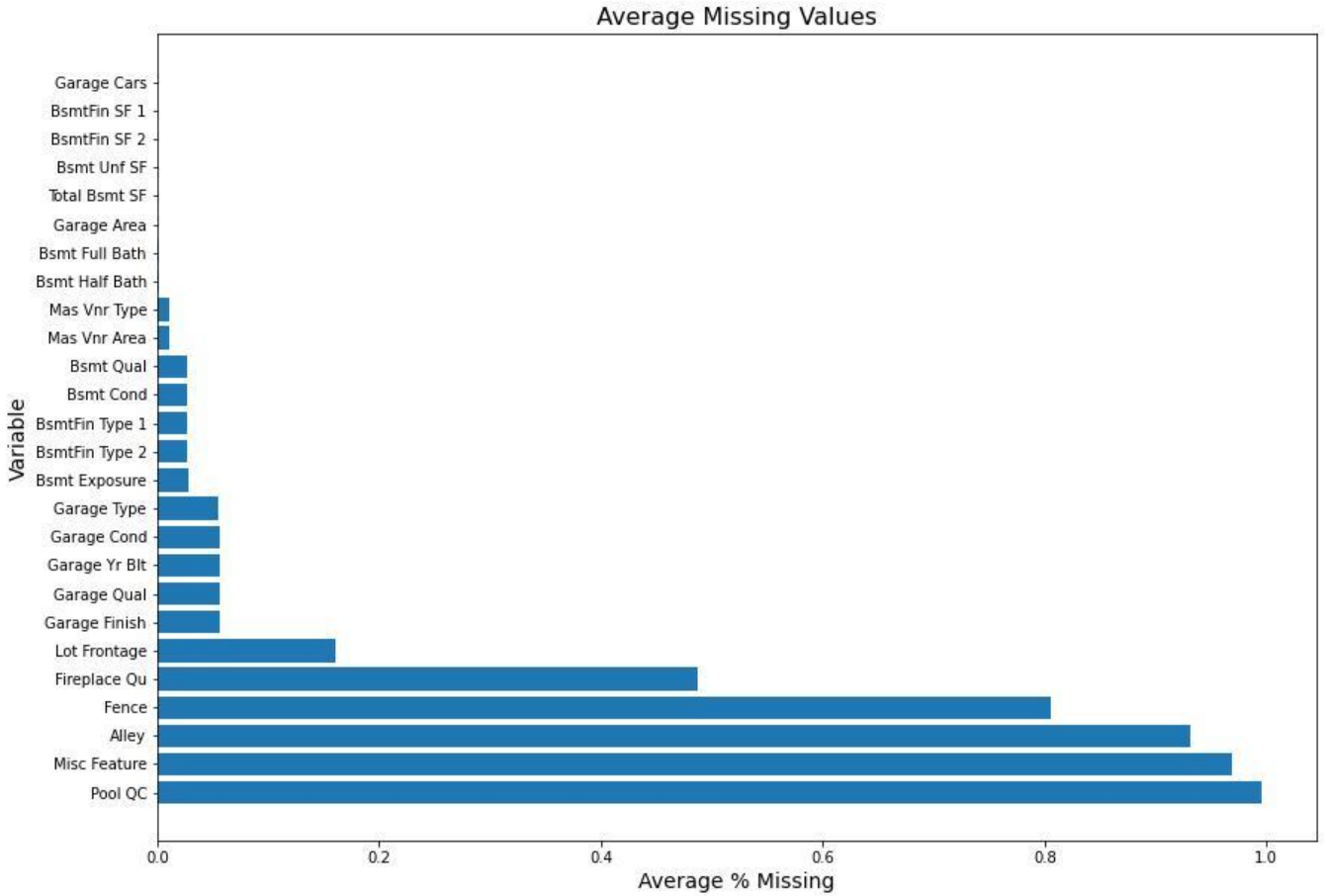
Prem Patel

Why Ames?

- Recently contracted by Iowa's Dept. of Housing & Urban Development
- Teach / consult their team how to build a good predictive model
- Cover methodology and workflow process I utilized for a ***preliminary*** model
- Necessary improvements

- **Problem:** lots of missing data

- **Solution:** dropping, or imputing



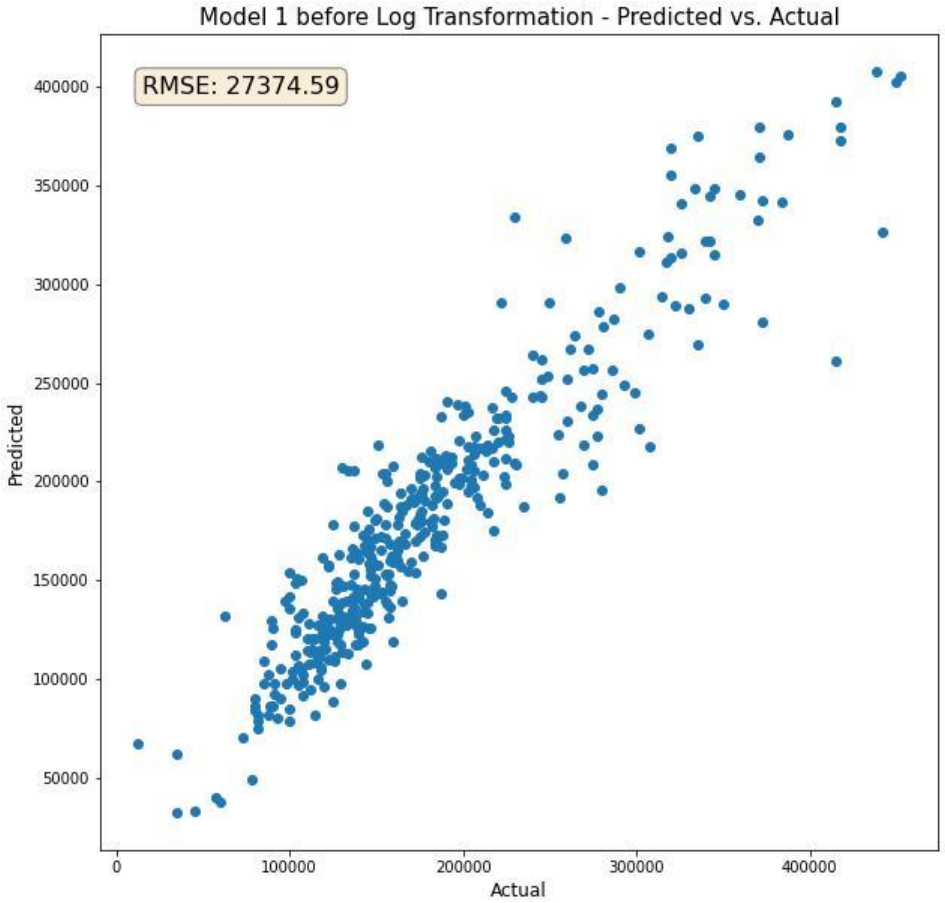
Which variables can help us predict?

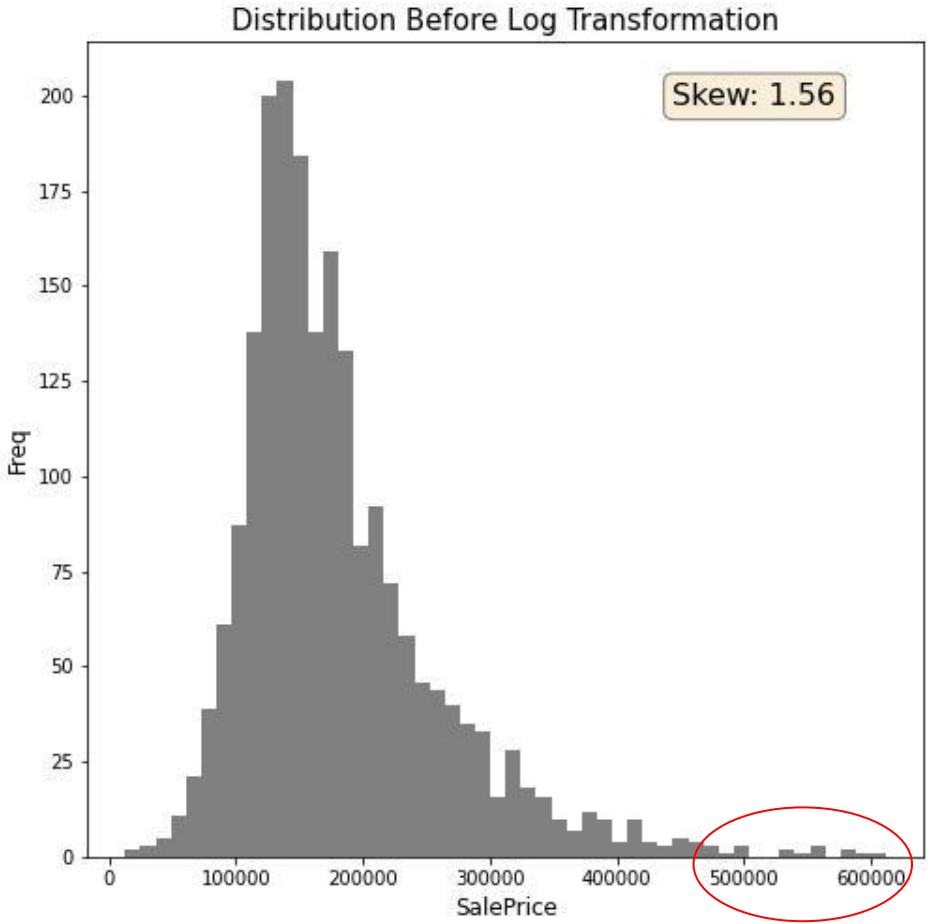
- Observing correlation between the variables of interest and our outcome of interest - Sale Price
- Ensuring all categorical dummies are included
- Creating new variables

	SalePrice
Overall Qual	0.800207
Gr Liv Area	0.697038
Garage Area	0.650246
Garage Cars	0.648197
Total Bsmt SF	0.628754
1st Flr SF	0.618486
Full Bath	0.537969
Foundation_PConc	0.529047
TotRms AbvGrd	0.504014
Mas Vnr Area	0.503579
Bsmt Qual_TA	-0.505320
garage_age	-0.516738
Garage Finish_Unf	-0.532545
Kitchen Qual_TA	-0.540860
remodel_age	-0.550370
age	-0.571849
Exter Qual_TA	-0.600362

My First Model

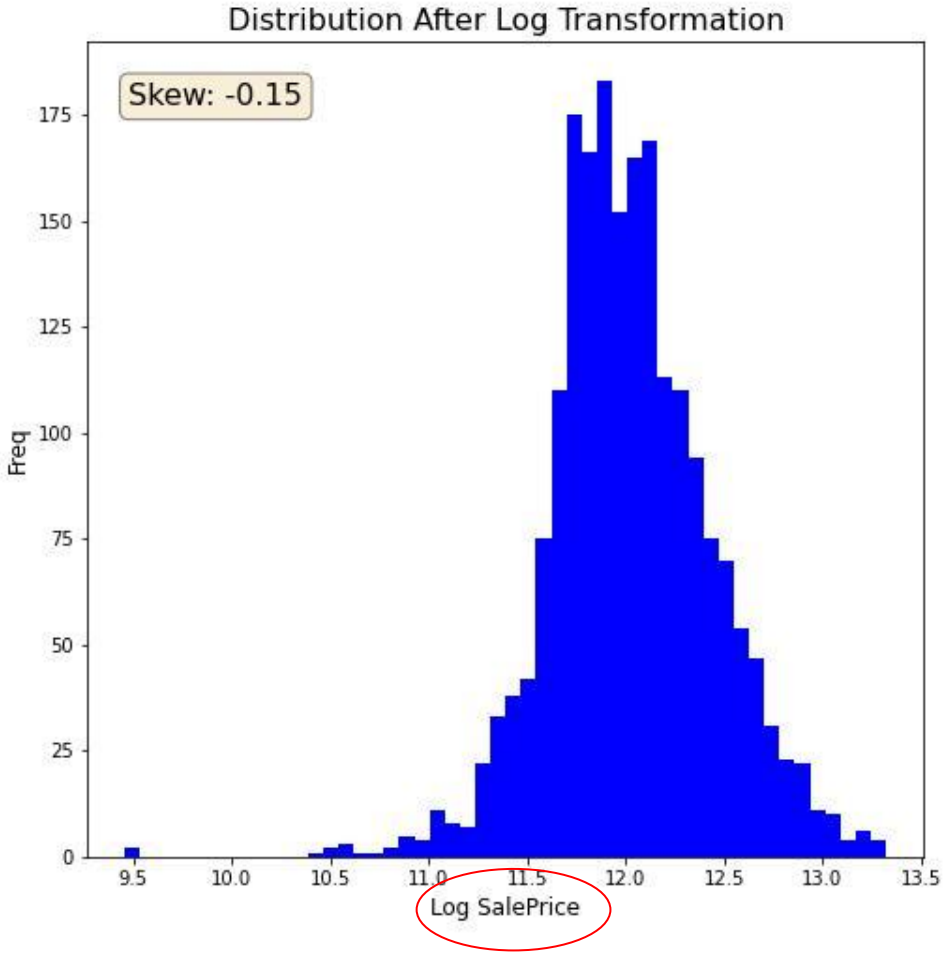
- Increased error for home prices greater than ~ $275,000





Is Sale Price Normally Distributed?

- Heavy right skew, likely being affected by outliers
- |Skew| > 1

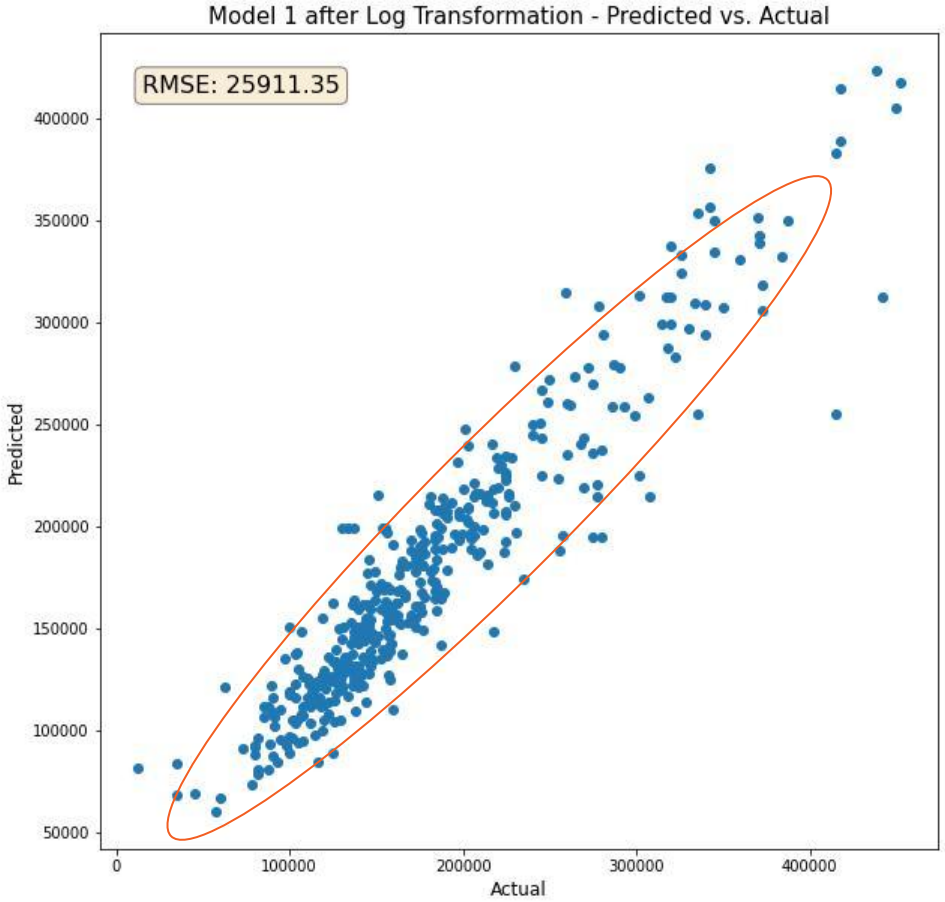


Is Sale Price Normally Distributed?

- Log Transforming the target variable
- Reduces skew
- Slightly more normalized
- Kurtosis

First Model after Log Transformation

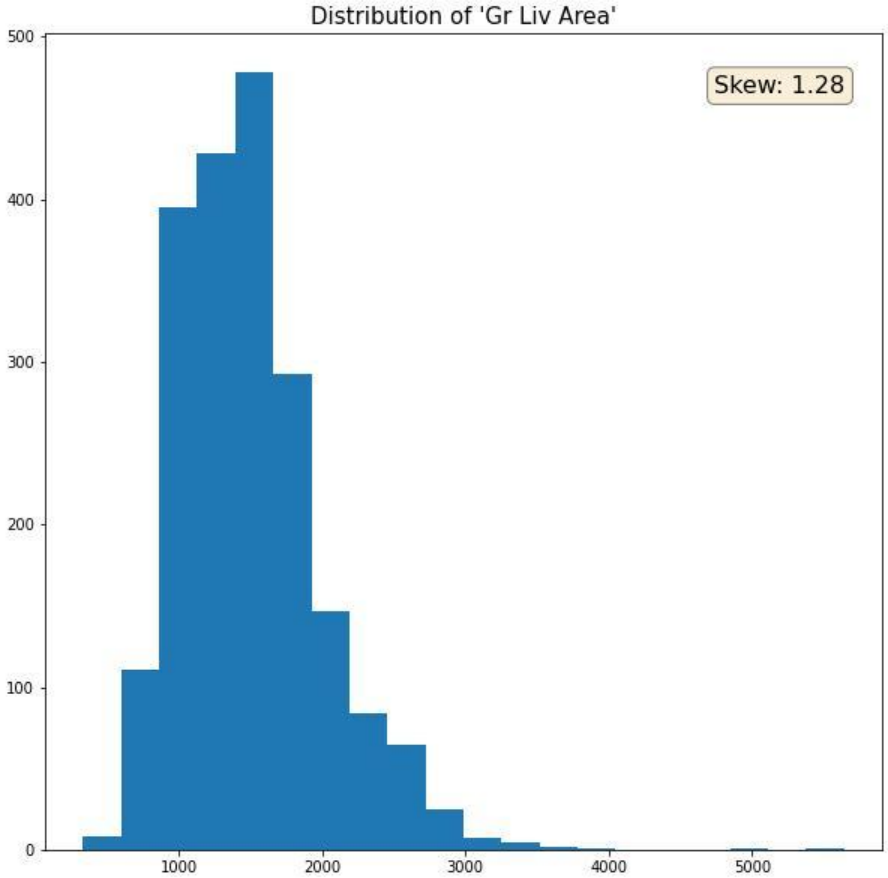
- Slightly tighter Prediction vs. Actual scatter plot
- Overall RMSE has been reduced
- Substantial improvement for lower priced homes



Ridge and Lasso

- Didn’t go improve predictive power as expected …

- Takeaways
  - Reduced feature set via Lasso
  - Trimming outliers



Highly Skewed Features

	Skew
Gr Liv Area	1.281492
1st Flr SF	1.635146
Total Bsmt SF	1.389436
Foundation_Wood	31.999977
Foundation_Stone	20.194030
Bsmt Qual_Po	45.287967
Foundation_Slab	7.577887
Bsmt Qual_Fa	5.590991
Kitchen Qual_Fa	6.381325
Exter Qual_Fa	8.718292

Final Model & Comparisons

	Training r2	RMSE
MLR Log Model 3	0.8633	24586.62
MLR Log Model 2	0.8369	25896.96
MLR Log Transform	0.8369	25911.35
Lasso 2	0.8365	26062.27
Lasso 1	0.8365	26063.20
Ridge	0.8357	26559.09
MLR Model 1	0.8270	27374.59

Improvements to Consider

- Detailed hyperparameter optimization
- Fine tuning the feature selection process
- Experimenting with different models

Thank You