**Problem Statement:** We will use the Ames dataset from the Kaggle House Prices competition to accurately predict house prices in Ames, Iowa. The dataset contains 79 predictive features and one response, sale price.

**Critiques:** We found 2 notebooks on Kaggle that we thought had some issues that we could fix to create a better solution. We describe these issues below.

Notebook 1:

<https://www.kaggle.com/rohitmathur100/housing-regression>

* In this user’s data preprocessing stage, they replace NAs in identified categorical features with the mode of those features. This is not wise and both skews the data towards the mode and effectively removes information from the dataset, because it fails to consider that “NA” is a valid class for most of the categorical features. NA for several of the features stands for a particular class, such as “No Alley”. When the dataset is imported into R, those text “NA”s are transformed into R NAs, and must be correctly accounted for.

Notebook 2:

<https://www.kaggle.com/tfrmarques/linearregression-pca-w-o-outliers>

* We thought that this user’s code was quite good overall. The data preprocessing did not have the same flaws as the first notebook. However, this user chose to use a linear regression with pca as their model. We believe that there is no compelling reason to use PCA for this problem, as despite the large number of features, we also have thousands of rows of data. This means that we should not have to worry too much about the curse of dimensionality. PCA also makes the final model significantly less interpretable, as it makes the final most important and least important variables uninterpretable principal components, instead of meaningful variables like living area above ground and basement size. Creating a model without using PCA allows us to determine these most and least important variables, providing further insight to end users of our model.