Ames Housing step 1:

Process: I selected the features for my multiple regression model based on my intuition of what would be the best predictors of sale price. Then, I split the data into 20/80 train-test datasets using a random\_state to ensure that the results were replicable and ran the model.

Results: The model had an r-squared score of 0.7482380948292935 meaning that the model is about 75% better at predicting the actual results than if we simply used the mean SalePrice. This is a relatively good score and especially when considering my non-scientific selection of features.

Learnings:

* Even with the relatively good r-squared score, the MSE is quite large indicating that there are some outliers in the dataset. Since the total MSE term is extremely large while the r-squared score is relatively good, it is reasonable to assume that there are a few outliers where the predictions significantly differ from the actual ‘SalePrice’.
* The features “Overall Qual” and “Overall Cond” are very large influences on the model with everything else being less predictive. Since Overall Qual and Overall Cond are bounded terms, the values in those columns are defined and not infinite, this makes us doubly sure that they are good predictors of SalePrice.
* Pool area is a negative coefficient, meaning the model believes a larger pool is associated with a decrease in SalePrice. This intuitively isn’t correct; the implication is that only houses with pools get compared here because houses with pools should have a larger SalePrice.

Next steps:

* Feature engineering to achieve 3 goals
  1. Handle outliers, possibly using z-scores or boxplots to determine if a percentile cap needs to be put on the data.
  2. Bound or standardize all features so they can be measured similarly to Overall Cond and Overall Qual
  3. Determine how Pool size influences the model and determine how to make having or not having a pool more predictive. A possibility is to create a binary feature to determine if a house has a pool or doesn’t; a similar feature could be created for basement and garage for example.
* More scientific selection of features. Move on from gut feel to a data validation process such as RFE where the feature that is least predictive is systematically eliminated, allowing for a more objective feature selection process.
* Determine correlation between features using VIF. If two features are correlated, Overall Cond and Overall Qual would probably be correlated, their effect to the model is gets outsize influence.