House Prices: Advanced Regression Techniques

1. Data Wrangling and EDA

- remove features with very high % of missing values (PoolQC, MiscFeature, Alley, Fence)

- define threshold for the correlation of numeric variables with SalePrice and select only the variables with an abs(correlation) > threshold

- (carefully) remove outliers

2. Feature Engineering

- log transform SalePrice

- impute remaining missing values (median for numeric values; define strategies for others)

- encode categorical and ordinal variables

- standardize predictors

- create new variables (e.g., TotalSF) and check correlation with SalePrice

- select best predictors

- remove features, impute, encode and standardize the test set after selecting the best predictors by reproducing what was done before *(this can also be done at the same time it is done for the training set as long as the two sets are dealt with **separately**)*

3. Modeling

- Linear Regression

- Decision Trees

- Random Forests

- XGBoost

- evaluate models with cross-validation