Notes on the data

Feature Selection

- "Alley" column can be ignored as there are a large number of Nan values

You can't use pd.get_dummies to one_hot enocode the data set. This is because the training set may not have the same number of types for a specific feature as the validation/test set. This will result in the model fitting more/less Columns than are expected by the model.predict and returns an error.

Notes on submissions

'forest_model_no-objects.csv'

In this submission we used no object features to train the model and very few salient numerical features. The features were:
['OverallCond', 'YearBuilt', 'OverallQual', 'LotArea', 'BedroomAbvGr']
No imputation was used.

'forest_model_no_obj_2.csv'

In this minimal submission we used only features which you would find on a website listing for a property.
['YearBuilt', 'OverallQual', 'LotArea', 'BedroomAbvGr', 'GarageCars']

Preprocessing

- The nan values for 'GarageCars' are replaced with 0.
- The data type of the train and test set did not match (int and float) for 'GarageCars'. The test set data was converted to float.

'forest_model_no_obj_3'

Numerical only model with lots of features.