1. Data Preprocessing

* Data is split into training and testing set, 20% for testing
* “ID” is dropped as it is only present during data scraping
* “Post date” feature is split into year and quarter of the date
* “Address”, “city”, “law document” are encoded into non-negative integers

1. Workflow

* GridSearchCV is first used for model selection and fitted on the entire training set.
* Then, the best performing model selected will be validated using K-Fold, with K = 5, and several metrics will be recorded: mean squared error (MSE), rooted mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE).
* Focus more on MAE and MAPE to conclude whether a model predicts well enough
* Feature importances are viewed through parameters of the model
* Repeat previous steps with the training set standardized and compare results to see whether the model needs data standardization or not

1. Models
   1. Linear models: OLS Linear Regression, Ridge Regression, Lasso Regression

* OLS and Ridge provide almost similar results, both with and without data standardization: MAE ~ 8.21, MAPE ~ 3.14
* Lasso with standardization has slightly lower errors: MAE ~ 8.18, MAPE ~ 3.13
* Pattern in feature coefficients:
  + property type, bedroom num., post year, and regional features (city and latitude+longitude) contribute the most, especially city and latitude, signaling a noticeable price difference in real estates at HN and HCMC
  + area has small value of coefficient, only noticeable with standardized data. This may be due to areas having large values while prices measured in billion VND are smaller.
  1. SVM Regressor
* Linear SVM: MAE ~ 9.52, MAPE ~ 4.07 -> worse than linear models; but with data standardization: MAE ~ 7, MAPE ~ 1.58 -> better
* Important features: Property type, num. of bedrooms, num. of bathrooms, year posted
* Non-linear SVM: (not yet)
  1. Random Forest Regressor
  2. Gradient Boosting
* Slightly higher MAPE (1.92) compared to Linear SVM (1.58), but lower MAE (5.87 vs 7) -> generalize better between lower-priced and higher-priced estates
* Important features: area, property type and bedroom num.
* With standardized data: MAE increase to 6.26, MAPE decrease to 1.84; property type, bedroom num. and bathroom num. have increased importance, while area drops from 0.274 to 0.026
* Possible reason in feature importance changes: estates with larger area should have higher prices -> with standardized values, differentiating prices based on area should be harder, whereas discrete variables like bedroom num. will still be discrete and therefore will have higher importance (in short, should not standardize data for tree-based models)