Notes of data\_wrangling\_tidyverse

# Data Wrangling Process

Data from <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

Data was already split into test and train data sets. - Train data set contained 1,460 rows and 81 columns - Test data set contained 1,459 rows and 80 columns (minus the SalePrice column, which will be predicted)

## Step 1: The test and train data sets were combined for pre-processing. How?

* A new column named ‘label\_split’ was added to both data sets to facilitate later re-split. The value ‘train’ was added to the ‘label\_split’ for train data set and the value ‘test’ for the test data set.
* The column ‘SalePrice’ was added to the test data set and the value NA was assigned to all rows.
* A merged data set was created using rbind() and contained 2,919 rows and 82 columns.

## Step 2: