Reference: https://www.kaggle.com/alexisbcook/introduction

- tackle data types often found in real-world datasets (**missing values**, **categorical variables**),
- design **pipelines** to improve the quality of your machine learning code,
- use advanced techniques for model validation (**cross-validation**),
- build state-of-the-art models that are widely used to win Kaggle competitions (**XGBoost**), and
- avoid common and important data science mistakes (**leakage**).
- Work on Housing Prices Competition for Kaggle Learn Users

Improve RandomForestRegressor

- n_estimators
- Criterion
- Random_state

Most machine learning libraries (including scikit-learn) give an error if you try to build a model using data with missing values.

Handling missing value feature

1. Drop columns with missing values
    - Good points: it drop missing values
    - Bad points: consider if this is important column and only one data row is missing, it will drop the whole column

2. Imputation
    - Fills in the missing values with some number; normally is the mean value along each column
    - Good thing: we preserve important column
    - Bad thing: not always accurate and imputed value systematically above or below the actual data that was not collected in the dataset

3. An extension to imputation
    - How ?
        i. Calculate same imputation value
        ii. Add a new column that indicate if the this row is imputed value or not