SUPER DATASCIENCE
MAKING THE COMPLEX SIMPL

CASE STUDY 022
[Python]
Predicting House Prices

Difficulty Level: 3 of 3

In this case study, we will use one Kaggle's **House Prices: Advanced Regression Techniques** dataset to predict house price prediction. We will only use train dataset, since we don't have the house price columns on test data.
Datasets can be found at: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

You are a data scientist helping a real estate industry to predict house prices. By doing this challenge, you will step into Machine Leaning World. House price prediction is very well known a common problem. During this case study, you will get familiar with one basic Machine Learning Algorithm, Linear Regression and one advance Machine Learning Algorithm, Random Forest.

Your analysis must be able to address the following requests:

1- Prepare your dataset to do prediction house prices
2- When your dataset is ready (no missing values, correct datatype, no unnecessary columns), start with Linear Regression.
3- After that, try to improve the prediction with regularization.
4- Try to predict the prices, using Random Forest
5- Show which features are the most helpful to predict house prices using random forest
6- In all steps, please cross validate your score

Acknowledgments

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset [1]

Good luck!
Difficulty note: this is a difficult assignment. Do not be surprised that there will be lots of nuances we have not covered off in the courses. But just like in the Real Life – there will be things training has not prepared you for and you will need to do research to find how to solve the problems at hand. If you get stuck, check the clues file.