SUPER DATASCIENCE
MAKING THE COMPLEX SIMPL

CASE STUDY 019
[Python]
Kaggle House Prices: Advanced Regression Techniques Competition

Difficulty Level: 2 of 3

In this case study, we will use one Kaggle's **House Prices: Advanced Regression Techniques** dataset to learn combine datasets, clean, do some exploratory data analysis (EDA) and prepare the dataset for Machine Learning Modeling. There will be 2 datasets from Kaggle; train and test datasets. In this study we will utilize both of the datasets.

The reason why they have train and test is because, Machine Learning requires both. But in our case, we will only do data cleaning and EDA. To start with, we will need to download the datasets and combine them using Python's package; pandas.

Datasets can be found at: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

You are a data scientist helping understand the data. By doing this challenge, you will learn how to combine datasets (there are different ways to do it, and we will learn how to decide which one works for us), data cleaning (data preparation) and EDA.

Your analysis must be able to address the following requests:

1- Load and Combine two datasets
2- Identify the missing values and replace them if necessary, and drop them
3- Cleaning the data. In this case, for some columns you can fill missing data.
4- Find out the minimum, maximum and average Lot Area.
5- Plot the correlation between the columns
6- Plot the distribution of LotArea, and do the same after normalizing the LotArea
7- Pair plot the following columns : 'LotFrontage','BsmtFinSF1','BsmtFinSF2','BsmtUnfSF'

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.