**Course-End Project: Feature Engineering**

**Abdullah Shah**

PGP Data Science – Purdue University

PC DS – Applied Data Science with Python

Source Code Jupyter Notebook structure:

* Importing Necessary Libraries
* Understand the dataset
* Separating categorical and numerical data
* EDA of categorical variables
  + Removing columns that have more than 75% NULL/NaN values
  + Removing observations containing Null/NaN any values
  + Replacing NULL/NaN values with mode of their respective columns, if NULL/NaN value %age <= 20 in that column, filling the rest with 0 as the missing value percentage is more than 20%, so thought filling with a constant makes more sense
  + Visualization 1: CountPlot for all categorical variables
  + Visualization 2: BoxPlot for all categorical variables against SalePrice (Output Variable)
  + Hypothesis Testing - Dropping categorical columns having p-value >= 0.05 against SalePrice(Output Variable)
* EDA of Numerical Variables
  + Removing columns that have more than 75% NULL/NaN values
  + Dropping columns where 0 value count is more than 90%
  + Filling missing values with mean()
  + Replacing <=0 values with a very small constant value {0.000001}, in order to apply a BoxCox transformation on them for normalization
  + Applying BoxCox transformation over columns having 0.5 < skew() < -0.6
  + Visualization 3: DisPlot for all numerical variables to visualize skewness
  + Visualization 4: Histogram for numerical\_data
  + Visualization 5: Histogram for all numerical variables to visualize distributions
  + Visualization 6: Pair Plot of distribution and density, plotted for four columns at a time because it was taking too long to plot for the whole dataset at once
  + Visualization 7: Correlation matrix for numerical\_data to check for significant variables
  + Visualization 8: Heatmap for the above correlation matrix
  + Delete one of the two columns if coef. of correlation >= 0.8
* Combining all the significant num. and cat. Variables
  + Outlier treatment for numerical features using Inter-Quartile Range (IQR)
  + Visualization 9: BoxPlot for all significant variables