# Assignment 4&5 - PySpark Preprosessing and Clustering

**Submission Deadline: *26th May 2021***

***Submit the code and a Report in the form of a word document on google classroom.***

## Assignment 4

***Question: PREPROCESS THE DATASETS given along with this assignment***

i. Identify the type of each data attribute. Indicate if the type of attribute is Nominal, Ordinal, Interval or Ratio

ii. Load the given dataset into ***Spark Data-Frames*** and perform the following preprocessing on it.

   a) Handle missing values
   b) Identify if an attribute has outliers or noise
   c) Apply measures of the central tendency and dispersion to **analyze numeric attributes**. That is, compute the mean, median, mode, range, variance,correlation for the attribute. Don't just give values explain analyze them.
   d) Would you apply preprocessing techniques like discretization or normalization on any attribute? Explain your answer. If yes, then apply the technique and share the results.

## Assignment 5

***Question: CLUSTER THE DATA using the PySpark built-in K-means clustering algorithm (this is provided in the Spark Library).***

a) Cluster the Movies dataSet using atmost **three** attributes to avoid curse of dimentionality. You can select the attribute based on the preprocessing.
b) Cluster the DataSet 2 for different value of K

*Run your algorithm for various values of K and different values of convergence, show the results in your report.*

*Use measures such as SSE(the sum of square error), silhouette co-efficient, and NMI (normalized mutual Index) to analyze the clustering results.*

*See Scikit for more info on the above measures*
*https://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html*