**FOUNDATIONS OF DATA SCIENCE**

**LAB ASSIGNMENT - 1**

**DATA PREPROCESSI\NG TECHNIQUES**

You will be working with responses from the **Stack Overflow Annual Developer’s Survey 2020** and performing various data preprocessing techniques on them. The dataset contains professional information about developers all around the world who are part of the Stack Overflow community. For more information, check out the schema file which gives you information about every column.

Link to Dataset - [Stack Overflow Annual Developer Survey](https://insights.stackoverflow.com/survey/)

After downloading the .csv file, load its contents into a dataframe on a Python notebook and **drop all columns except the following** - Respondent, MainBranch, Hobbyist, Age, Age1stCode, CompFreq, CompTotal, ConvertedComp, Country, CurrencySymbol, DevType, Employment, Gender, NEWDevOps, NEWOtherComms, OpSys, SOAccount, SOComm, SOVisitFreq, SurveyEase, UndergradMajor, WorkWeekHrs, YearsCode, YearsCodePro.

Now, perform the following operations:

1. Rename some of the headers in the dataframe:
   1. NEWDevOps - DevOpsPresence
   2. NEWOtherComms - OtherComms
   3. OpSys - OS
2. Display the first and last 15 rows of the data frame.
3. Display information about the dataframe such as datatypes of all columns and so on. Convert the datatypes of columns to the field’s appropriate format. For example, years must be int, Undergrad major must be string etc.
4. Provide a statistical summary of each column e.g. count, column mean value, column standard deviation, etc.
5. Count the number of missing values in the columns – Age, Country, UndergradMajor
6. To deal with the missing data, perform the below-mentioned operations:
   1. For column ‘WorkWeekHrs’, replace the missing values by mean
   2. For column ‘SurveyEase’, replace the missing value by frequency
   3. For column ‘WorkWeekHrs’, perform binning with the following bins:
      * Low
      * Normal
      * High

Also, visualize the bins through a histogram.

1. In order to perform data standardization, filter the dataframe to include only ‘USD’ and ‘INR’ in the ‘CurrencySymbol’ column. Now ensure that all values in ‘ConvertedComp’ are in INR by converting from USD to INR where necessary.
2. In the column ‘Hobbyist’, introduce an indicator variable by assigning 0 for ‘No’, 1 for ‘Yes’ and -1 for missing values.