FOUNDATIONS OF DATA SCIENCE (CSD355)
LAB ASSIGNMENT - 2
DATA PREPROCESSI\NG TECHNIQUES
(Deadline: 04-09-2025)

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

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:
    a) NEWDevOps - DevOpsPresence
    b) NEWOtherComms - OtherComms
    c) 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:
    a) For column 'WorkWeekHrs', replace the missing values by mean
    b) For column 'SurveyEase', replace the missing value by frequency
    c) For column 'WorkWeekHrs', perform binning with the following bins:
        - Low
        - Normal
        - High

7) 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. (1 USD = 87 INR)
8) Convert the Gender column from object type to numeric type. Consider all the different gender categories in the dataset and assign each category a numeric value. Create a new column named Gender_numeric with these mapped values and add it to the dataframe.
9) Print the coding experience of the top 10 countries, considering the country-wise average coding experience.
10) The DevType column allows multiple roles separated by ;, split it and count how many respondents identify as each role (e.g., Web Developer, Data Scientist, etc.).