Name: Aditya Wanjale
Roll no: 31282

PICT

DSBDAL Assignment No-02

- **TITLE:** Data Wranggling II

- **PROBLEM STATEMENT:**

Create an academic performance dataset of students and perform the following operations using python.

1. Scan all Variables for missing values and inconsistencies. If there are any missing values and/or inconsistencies, use any of the suitable techniques to deal with them.
2. Scan all the numeric values for outliers, use suitable techiques to deal with them.
3. Apply data transformation on atleast one of the variable

- **LEARNING OBJECTIVES:**

1. Implement data preprocessing techniques on raw data.
2. Use python libraries to handle inconsistencies and irregularities in data.

- **LEARNING OUTCOMES:**

1. Students should be able to handle irregularities and inconsistencies in the raw, unformatted data, using python.

- **THEORY:**

Data wrangling is the process of gathering, collecting and transforming raw data into another format for better understanding, decision making and analysis in less time. Data wrangling deals with the following functionalities:-

1. Data exploration - In this process, the data is studied, analyzed and understood by visualizing representing of data.

2. Dealing with missing value - Most of the datasets have missing values. They have to be replaced or the data entry has to be dropped.
3. Reshaping the data - Adding or modifying data according to the requirements.
4. Filtering data - Removing unwanted rows or columns from the dataset

- Libraries and functions used:-
1. Pandas - For data exploration and visualization
   a) pd.readcsv("data.csv") - Load .csv file into dataframe
   b) df.describe() - gives columnwise details for numeric values.
   c) df['col-name'].fillna(value) - fill all missing values with specified value.
2. Matplotlib.pyplot - For data visualization
   a) plt.hist(column_name) - draw/plot the histogram for values in the column.

   b) plt.subplot() - multiple graphs within 1 figure.

3. Scikitlearn - For data tranformation
   a) Label_encoder(col_name) - Assign unique labels to each values in column. Only applicable to object/categorical column types.

- CONCLUSION :

Hence we applied data preprocessing techniques for unformatted, raw data and handled outliers, missing values and inconsistent entries. We also used data transformation techniques on the databases.