CS 6103D Software Systems Lab
Python Programming

Evaluation#1 : 06.12.2021 **Time: 45 Mins** **Maximum Marks: 6**

Q. The Student_marks.csv file consists of the marks secured by the students in a locality for an aptitude test concentrated on three main areas, viz., math, reading and writing. The given dataset contains the details of 1000 students. Read the file in python and perform the following operations:

1. Read the data present in the file into a dataframe named as 'student' by considering the first raw in the csv file as its header. There are noises in the math, reading and writing score attributes of the dataset. In particular, there are a few NaN values and missing values in them. For each of the three given attributes in the dataframe find out the noises separately and update the student dataframe by replacing the noises with the mean score value of the corresponding attribute. After removing the noise, display the mean and standard deviation of math, reading and writing scores. **(1.5 Marks)**

2. Note that each row in the dataframe represents the details of a unique student. Create a Python dictionary to store the marks of each student using the following logic:

   Create a unique ID for each of the students starting with "S", followed by the student number which is the corresponding raw number in the dataset, and finally attach the student's group to the ID using "_". For example, for a student present in the 123rd raw of the dataset with the group as "group B", his/her unique ID will be "S123_B".

   The key of the dictionary should be the unique ID of the student, and the value of the dictionary should be a list containing the student's scores for math, reading and writing. The dictionary would look like the following:

   ```
   {
           S1_B : [72,72,74 ],
           S2_C : [69,90,88 ],…
   }
   ```

   Take the ID of any student as input from the user and print the corresponding dictionary entry. **(1.5 Marks)**

3. Write a function named as **student_stat()** that takes in the student dataframe as its argument and returns a new dataframe called 'table' containing the following statistics: The table dataframe must contain the mean and standard deviation of the scores obtained by the students in the math section for the five different groups (A,B,C,D,E). The last column in the table calculates the probability for a student in each of the group to score **more than** 50 marks in Math. The table should have the following format:

Group	Mean math score	Stan. Dev. math score	Prob_scoring above 50_maths
A			
B			
C			
D			
E			

(3 Marks)