VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Digital Assignment - II

Programme	:	**B.Tech (CSE)**	Semester	:	**WIN:2020-2021**
Course	:	**Large Scale Data Processing**	Code	:	**CSE3025**
Faculty	:	**Dr. Ramesh Ragala**	Slot	:	**G2**
Class Number	:	**CH2020215001362**		:	

Slove the following assignment questions programmatically. Write your reg.no and name on every sheet of the answer script.

1. In this Assignment you need to understand the concept of MapReduce using python. For simplicity, you need to work on snapshot of a large dataset made of (index value) couples separated by a newline, which is shown in Table - 1.

Index	**Variable - 1**
0	11839923.64831265
1	5710431.90800272
2	3782569.12679777
3	15897765.23578973
4	13609375.26506385

Table 1: Snapshot of large dataset

   (a) Develop a Python based MapReduce program to compute minimum value, maximum value, average value, variance value, standard deviation, median, and mode of variable - 1 of attribute in above dataset.

   (b) Develop a Python based MapReduce program to perform normalization techniues such as rescaling, mean normalization, z-score normalization and Standardization techniques on variable - 1 of the dataset.
   Hint: refer the following
   1. https://developers.google.com/machine-learning/data-prep/transform/normalization
   2. https://towardsdatascience.com/data-normalization-in-machine-learning-395fdec69d02
   3. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

2. Develop a Python based MapReduce program to perform Matrix multiplication of two matrices, one of the matrix with a dimension as m $\times$ n and the other matrix with dimension as n $\times$ p.
   Hint: refer the following
   1. https://eis.hu.edu.jo/deanshipfiles/conf112262685.pdf