Course: **Data Mining** 23-Jan-2023 **(Fall 2022)**

Resource Person: **Dr. Muhammad Faheem**

ASSIGNMENT–2 (Introduction and Getting to know Data)

Total Points: 20
Submission Due: Saturday Jan 27, 2023
(Google Classroom Course Page)

Instructions: Please Read Carefully!

- **This is an individual assignment. Everyone is expected to complete the given assignment on their own, without seeking any help from any website or any other individual.** There will be strict penalties for any work found copied from any source and the university policy on plagiarism will be strictly enforced.
- Assignment is to be submitted via Google Classroom.
    - You should already have created your account on Google Classroom as per my earlier email. If not, then follow the link in that email to create your account.
    - Submit your assignment on or before due date. **No late submissions will be possible**.
    - The viva of this assignment will be conducted.

Question 1: **[5]**

Given two objects represented by the tuples (42, 5, 22, 15) and (35, 1, 15, 13):

(a) Compute the *Euclidean distance* between the two objects.
(b) Compute the *Manhattan distance* between the two objects.
(c) Compute the *supremum distance* between the two objects.

Question 2: **[5]**

Explain with examples the following concepts in Data Integration:

(a) Entity Identification Problem
(b) Redundancy and Correlation Analysis for Nominal and Numeric data

Question 3: **[10]**

Consider the data for the attribute marks: 33, 35, 36, 36, 39, 40, 40, 41, 42, 42, 45, 45, 45, 45, 50, 53, 53, 55, 55, 55, 55, 56, 60, 65, 66, 72, 90.

(a) Use smoothing by bin means to smooth these data, using a bin depth of 3.

(b) How might you determine outliers in the data?

(c) Use min-max normalization to transform the value 55 for marks onto the range [0.0, 1.0].

(d) Use z-score normalization to transform the value 35 for marks.

(e) Plot an equal-width histogram of width 10.

(f) Sketch examples of each of the following sampling techniques: SRSWOR, SRSWR, cluster sampling, and stratified sampling. Use samples of size 5 and the strata “Below-Average”, “Average” and “Exceptional”. Compute an approximate median value for the data.