COMPUTER SCIENCE 4373/5473
Assignment #1

Points: 100 **Weight**: 2%
Due: Friday, Sept. 4, 2020 at 11:59pm on Blackboard
Note: Late assignment will not be accepted without instructor's pre–approval.

Instruction: In this assignment, you will work individually to develop a program for the pre-processing data. You will use Python in Anaconda distribution to write a Jupyter notebook. Your program will read data from the provided CSV file into a DataFrame. The data file has 6 columns: A, B, C, D, E, and F, where A and B are categorical and the rest are numeric. Your program will than solve the following problems and write output to a text file. All values in the output should be formatted to have up to four (4) digits after the decimal point. Unless explicitly stated otherwise, you should use packages such as pandas and numpy that come installed with Anaconda. Submit your solution via BlackBoard Learn in a zipped file yourName-hwk01.zip, which contains the notebook, additional library if any, and the input/output files.
Notice: *Given the nature of on-line course, we will require you to practice using Words, Markdown, or HTML to write and format your homework solutions (**no scanned smeared image please**). It will prepare you for taking the on-line exams, where only a Words style editor (with HTML support) is available.*

1. [**40**] (Data Statistics) Write basic Python functions to obtain the following statistics and apply these functions to columns C, D, E, and F in the table.
   (a) The mean and the midrange.
   (b) The mode and the modality (i.e., bimodal, trimodal, etc.).
   (c) The five-number summary.
   (d) Compare to the corresponding functions provided by DataFrame
2. [**30**] (Similarity and Distance) Prompt the user for a tuple, say $p = (a1, b2, 515, -0.876, 6.4253, 45)$, and perform the following tasks.
   (a) Print the row in the DataFrame that is the least dissimilar to p, where the dissimilarity is measured by different types of distances, including the Euclidean distance, Manhattan distance, supremum distance, and cosine similarity, using the set of columns C, D, E, and F.
   (b) Normalize the data points by making the norm of each data point (under columns C, D, E, and F) equal to 1. That is, scale the values in columns C, D, E, and F, so that, for each row (c, d, e, f) we have $\sqrt{c^2 + d^2 + e^2 + f^2} = 1$. Then, print the row in the DataFrame that has the shortest Euclidean distances from the normalized point p.
3. [**30**] (Normalization) Write functions to normalize the data in a given column using the following methods. Apply these functions on column C.

(a) Min-max normalization that transforms the values onto a given range, for example, $[-1.0, 1.0]$.

(b) Z-score normalization.

(c) Decimal scaling normalization.