COMPUTER SCIENCE 4373
Assignment #1

Points: 100 **Weight**: 4%
Due: Friday, January 26, 2018 at 11:55pm 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 may use Java in an Eclipse environment or Python in Anaconda distribution. Your program should reads data from a CSV file and write output to a text file. You will submit a zipped file yourname-hwk01.zip in BlackBoard Learn, which should contain the entire Java project or Python project (Jupyter notebook) including program source code, required library, and the input/output files.

Requirements. You should develop either a Java program or a Python (Jupyter) notebook. You should use Eclipse for Java program and Anaconda for Python notebook. You should use the provided input file to generate the output (some editing may be allowed on the output file for getting the complete answers). All values in the output should be formatted to have up to four (4) digits after the decimal point. Unless explicitly prohibited, you can take advantage of libraries that come installed with Weka or Anaconda.

Set up. Your program should allow the user to enter the name of the input file and load the input into a data table. This file will have 6 columns: A, B, C, D, E, and F, where A and B are categorical and the rest are numeric. The following questions assume you have this setup already. In the following, the words "function" is used to refer to Java method and Python function.

1. [**15**] (Data Statistics) Write 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.
2. [**15**] (Smoothing) Write functions to smooth data in a given column using the following methods. Apply these function on column F.
   (a) Equal-depth binning with bin means for depth k, for example $k = 100$.
   (b) Equal-depth binning with bin boundaries for depth k, for example $k = 100$.
   (c) Equal-width binning with bin median for 10 bins.
3. [**15**] (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.

4. [**10**] (Data Reduction) Write a function that takes a data table, a set of column names (of numeric columns), and an integer p (less than the total number of columns in the table), and use PCA method to reduce the set of columns into p new columns. Apply this function to reduce the columns {C, D, E, F} into two columns p1, and p2. (You may use the pca method in Weka or in Scikit-Learn).

5. [**20**] (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 to the row in the table that is most similar (least distance) from p based on values in columns C, D, E, and F, according to Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.

   (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 with the least Euclidean distances from the normalized point p.

6. [**25**] (Correlation)

   (a) Compute the covariance and the correlation coefficient for each pair of the numeric columns.

   (b) Compute the contingency table for columns A and B, similar to the following one. Use the Pearson's chi-square (χ^2) test of independence with a confidence level of 0.001 to determine if the two attributes are correlated. Notice that the distinct values in attribute A are $\{a_1, a_2\}$ and that in attribute B are $\{b_1, b_2, b_3\}$. Write the result to the output file.

		A		
		a_1	a_2	all
B	b_1	??	??	??
	b_2	??	??	??
	b_3	??	??	??
	all	??	??	??