CSCI 5080 Assignment 1

Use Microsoft **WORD** to write your answer and submit it to the Dropbox in D2L.
Please indicate **how much time** you spend on each problem.

1. (25 points) Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 11, 13, 15, 17, 19, 21, 21, 23, 23, 23, 23, 25, 27, 30, 33, 33, 33, 33, 36, 36, 38, 40, 46, 48, 54.

(a) What is the *mean* of the data? What is the *median*?

(b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).

(c) What is the *midrange* of the data?

(d) Can you find (roughly) the first quartile ($Q1$) and the third quartile ($Q3$) of the data?

(e) Give the *five-number summary* of the data.

(f) Show a *boxplot* of the data. (Watch the video on creating a simple boxplot in Excel.)

2. (25 points) Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result

age	20	22	25	25	36	40	45	48	49
%fat	8.4	25.3	7.6	18.8	27.5	24.6	28.1	28.8	30.2
age	51	53	53	57	58	59	60	61	62
%fat	32.7	40.2	29.8	32.3	30.7	33.9	40.1	33.1	36.4

(a) Calculate the mean, median and standard deviation of *age* and *%fat*.

(b) Give the *five-number summary* of the data.

(c) Draw the boxplots for *age* and *%fat* in Excel.

(d) Draw a *scatter plot* based on these two variables in Excel.

3. (20 points) Given two objects represented by the tuples (15, 7, 24, 21) and (12, 0, 16, 10):

(a) Compute the *Euclidean distance* between the two objects.

(b) Compute the *Manhattan distance* between the two objects.

(c) Compute the *Minkowski distance* between the two objects, using $h = 3$.

(d) Compute the *supremum distance* between the two objects.

4. (30 points) It is important to define or select similarity measures in data analysis. However, there is no commonly accepted subjective similarity measure. Results can vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation. Suppose we have the following two-dimensional data set:

	A_1	A_2
x_1	1.6	1.8
x_2	2.1	1.6
x_3	1.7	1.2
x_4	1.2	1.4
x_5	1.5	1.3

(a) Consider the data as two-dimensional data points. Given a new data point, $\boldsymbol{x}$ = (1.3, 1.5) as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.

(b) Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.

Hint: Use the following equations to normalize (A_1, A_2) to (A_1', A_2') so that $\sqrt{(A_1')^2 + (A_2')^2} = 1$

$$\left(A_1' = \frac{A_1}{\|x\|} = \frac{A_1}{\sqrt{(A_1)^2+(A_2)^2}} \quad , \; A_2' = \frac{A_2}{\|x\|} = \frac{A_2}{\sqrt{(A_1)^2+(A_2)^2}}\right)$$