CSE 4095 - Introduction to Data Science and Engineering	Spring 2024
Assignment 1	Due date: February 10th, 11:59pm, EST

For full credit, please adhere to the following:

- Unsupported answers receive no credit.
- All answers can be typed or handwritten, and should be readable.
- Submit the assignment in one file (.pdf, .doc, etc.) via HuskyCT.

The following questions are in the textbook for the course, *Data Mining: Concepts and Techniques (3rd Edition)*, by Jiawei Han, Micheline Kamber, Jian Pei.

1. (14 points) Exercise Set 2.2

2.2 Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(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 (Q_1) and the third quartile (Q_3) of the data?
(e) Give the *five-number summary* of the data.
(f) Show a *boxplot* of the data.
(g) How is a *quantile–quantile plot* different from a *quantile plot*?

2. (15 points) Exercise Set 2.4

2.4 Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Calculate the mean, median, and standard deviation of *age* and *%fat*.
(b) Draw the boxplots for *age* and *%fat*.
(c) Draw a *scatter plot* and a *q-q plot* based on these two variables.

3. (16 points) Exercise Set 2.6

2.6 Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

(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 $q = 3$.

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

4. (10 points) Exercise Set 2.8

2.8 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 2-D data set:

	A_1	A_2
$\boldsymbol{x_1}$	1.5	1.7
$\boldsymbol{x_2}$	2	1.9
$\boldsymbol{x_3}$	1.6	1.8
$\boldsymbol{x_4}$	1.2	1.5
$\boldsymbol{x_5}$	1.5	1.0

(a) Consider the data as 2-D data points. Given a new data point, $\boldsymbol{x} = (1.4, 1.6)$ 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.

5. (15 points) Exercise Set 3.3

3.3 Exercise 2.2 gave the following data (in increasing order) for the attribute *age*: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) Use *smoothing by bin means* to smooth these data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

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

(c) What other methods are there for *data smoothing*?

6. (16 points) Exercise Set 3.7

3.7 Using the data for *age* given in Exercise 3.3, answer the following:

(a) Use min-max normalization to transform the value 35 for *age* onto the range [0.0, 1.0].

(b) Use z-score normalization to transform the value 35 for *age*, where the standard deviation of *age* is 12.94 years.

(c) Use normalization by decimal scaling to transform the value 35 for *age*.

(d) Comment on which method you would prefer to use for the given data, giving reasons as to why.

7. (14 points) Exercise Set 3.8

3.8 Using the data for *age* and *body fat* given in Exercise 2.4, answer the following:

(a) Normalize the two attributes based on *z-score normalization*.

(b) Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two attributes positively or negatively correlated? Compute their covariance.