Professor:
Dr. Laura E. RAILEANU

Assignement Nº 2

Data Preprocessing

A.

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* (Q1) and *the third quartile* (Q3) of the data?

(e) Give *the five-number summary* of the data (the minimum value, first quartile, median value, third quartile, and maximum value)

(f) Show a *boxplot* of the data.

B.

In real-world data, tuples with *missing values* for some attributes are a common occurrence. Describe various methods for handling this problem.

C.

Using the data for *age* given in A, answer the following.

(a) Use *smoothing by bin means* to smooth the above 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*?

(a)

The following steps are required to smooth the above data using smoothing by bin means with a bin depth of 3.

Step 1: Sort the data. (This step is not required here as the data are already sorted.)

Step 2: Partition the data into equal-frequency bins of size3.

D.

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

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) Normalize the two variables based on *z-score normalization*.

(d) Calculate the *correlation coefficient* (Person's product moment coefficient). Are these two variables positively or negatively correlated?

E.

Using the data for *age* given in Exercice A,

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

(b) Sketch examples of each of the following sampling techniques: SRSWOR, SRSWR, cluster sampling, stratified sampling. Use samples of size 5 and the strata “youth", “middle-aged", and ”senior".