PROBLEM SET 2. DATA PREPARATION

P1. In real-world data, tuples with *missing values* for some attributes are a common occurrence. Describe various methods for handling this problem.

P2. Excercise 1 (in the Problem set 1) 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*?

P3. Discuss issues to consider during *data integration*.

P4. What are the value ranges of the following *normalization methods*?

(a) min-max normalization
(b) z-score normalization
(c) z-score normalization using the mean absolute deviation instead of standard deviation
(d) normalization by decimal scaling

P5. Using the data for *age* given in Exercise 1, 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.

P6. Propose an algorithm, in pseudo code or in your favorite programming language, for the following: The automatic generation of a concept hierarchy for nominal data based on the number of distinct values of attributes in the given schema.