CSE3212 Data Mining
Week 4 Tutorial
Data Preparation

1. Suppose that the data from analysis include attribute age. The age values from the tuples are 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
   a. What is the mean of the data?
   b. What is the median of the data?
   c. What is the mode of the data?
   d. Is the age has a skewed data distribution?
   e. Is there any outlier?

2. Data cleaning, data integration and transformation, data reduction and data discretization are activities that may be carried out during data preprocessing. Within each of activities there are a number of possible techniques.
   a. Would descriptive statistics be useful in helping you to decide the choice of techniques in all data preparation activities?
   b. Which activity will be benefited from descriptive statistics, give an example.

3. Using the age data in question 1, answer the following.
   a. Use the smoothing by bin means to smooth the data, using bin depth of 3.
   b. Which binning method would you consider best to use for the age data? Why?

4. Using the data 200, 300, 400, 600, 1000 answer the following.
   a. What type of normalization would you perform if the business requires this normalized data to fall within the range of 1-100? Perform the normalization for the data.
   b. What type of normalization would you perform if the business requires this normalized data to be distributed as close as possible to a normal distribution? Perform the normalization of the data.

5. Suppose a group of 12 sales price records has been sorted as follows:
   5,10,11,13,15,35,50,55,72,92,204,215
   Partition them into three bins by each of the following methods:
   a. Equal-frequency partitioning.
   b. Equal-width partitioning
   c. Clustering

   Which method would you choose to discretize the sales data? Why?