**DATA PREPROCESSING**

**Practice Questions:**

**Question1:**

Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

Example: Age in years. Answer: Discrete, quantitative, ratio

(a) Time in terms of AM or PM. Binary, qualitative, ordinal  
(b) Brightness as measured by a light meter. Continuous, quantitative, ratio  
(c) Brightness as measured by people’s judgments. Discrete, qualitative, ordinal  
(d) Angles as measured in degrees between 0◦ and 360◦. Continuous, quantitative, ratio  
(e) Bronze, Silver, and Gold medals as awarded at the Olympics. Discrete, qualitative, ordinal  
(f) Height above sea level. Continuous, quantitative, interval/ratio (depends on whether sea level is regarded as an arbitrary origin)  
(g) Number of patients in a hospital. Discrete, quantitative, ratio  
(h) ISBN numbers for books. (Look up the format on the Web.) Discrete, qualitative, nominal (ISBN numbers do have order information, though)

(i) Ability to pass light in terms of the following values: opaque, translucent, transparent. Discrete, qualitative, ordinal  
(j) Military rank. Discrete, qualitative, ordinal  
(k) Distance from the center of campus. Continuous, quantitative, interval/ratio (depends)  
(l) Density of a substance in grams per cubic centimeter. Discrete, quantitative, ratio  
(m) Coat check number. (When you attend an event, you can often give  
your coat to someone who, in turn, gives you a number that you can  
use to claim your coat when you leave.) Discrete, qualitative, nominal

**Question 2.** Can you think of a situation in which identification numbers would be useful for prediction?  
One example: Student IDs are a good predictor of graduation date.

Question 3:

Distinguish between noise and outliers. Be sure to consider the following  
questions.  
(a) Is noise ever interesting or desirable? Outliers?  
No, by definition. Yes.   
(b) Can noise objects be outliers?  
Yes. Random distortion of the data is often responsible for outliers.  
(c) Are noise objects always outliers?  
No. Random distortion can result in an object or value much like a normal one.  
(d) Are outliers always noise objects?  
No. Often outliers merely represent a class of objects that are different from normal objects.  
(e) Can noise make a typical value into an unusual one, or vice versa?  
Yes.

**Question 4:**

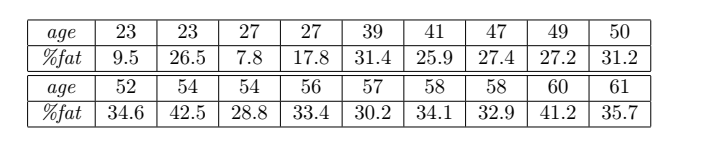
You are given a set of *m* objects that is divided into *K* groups, where the *ith* group is of size *mi*. If the goal is to obtain a sample of size *n < m*, what is the difference between the following two sampling schemes? (Assume sampling with replacement.)

(a) We randomly select *n ∗ mi/m* elements from each group.  
(b) We randomly select *n* elements from the data set, without regard for the group to which an object belongs.

Answer: The first scheme is guaranteed to get the same number of objects from each group, while for the second scheme, the number of objects from each group will vary. More specifically, the second scheme only guarantes that, on average, the number of objects from each group will be *n ∗ mi/m*.

**Question 5: For details on the box plot, see page 54 of the book “Datamining Concepts and Techniques”.**

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

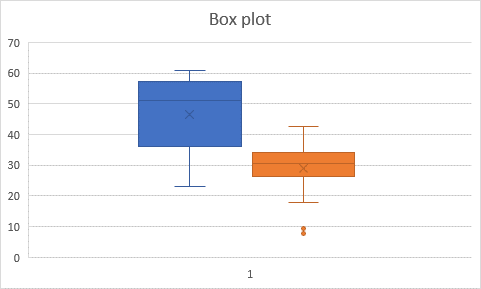


**ANSWER**

(a) Calculate the mean, median and standard deviation of *age* and *%fat*.  
For the variable *age* the mean is 46.44, the median is 51, and the standard deviation is 12.85.

For the variable *%fat* the mean is 28*.*78, the median is 30.7, and the standard deviation is 8.99.

(b) Draw the boxplots for *age* and *%fat ….*



**Scatter plot**