

## Computer Science COMP3420 in 2009 – Assignment One

**Due:** 5pm Thursday, April 9

**Late Penalty:** 20% per day

*No programming is needed for this assignment. Before you hand in your assignment, make sure you have put your name, student ID and your tut/lab group into the front page of your assignment. Put your work into the COMP3420 box on the ground floor in CSIT Building. The total mark for this assignment are 20 points (20% for the final course marks).*

**Question 1** (4/20) What are the differences between a data warehouse and an operational database? Can you describe the main components of a data warehouse and their functions?

**Question 2** (3/20) Suppose a group of 19 *sensor reading* records has been sorted as follows.

4, 5, 7, 8, 10, 11, 13, 14, 15, 20, 35, 40, 45, 50, 55, 62, 75, 92, 105.

Partition them into four bins using the following four methods.

- (a) equal-width partitioning
- (b) equal-frequency partitioning
- (c) smooth the data in bins by bin boundaries
- (d) smooth the data in bins by bin means.

**Question 3** (4/20) Suppose 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) Normalise the two variables *age* and *%fat* using *z-score normalization*.

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

**Question 4** (5/20)

Suppose that a data warehouse consists of the three dimensions *time*, *doctor*, and *patient*, and the two measures *count* and *charge*, where *count* is the number of visits of a patients to a doctor and *charge* is the fee that a doctor charges a patient for a visit.

- (a) Draw a *star* and a *snowflake* schema diagrams for the above data warehouse.
- (b) Starting with the base cuboid [day, doctor, patient], what specific *OLAP operations* (e.g. roll-up) should be performed in order to list the total fee collected by each doctor in 2008?
- (c) To obtain the same list as (b), write an SQL query assuming the data is stored in a relational database with the schema *fee*(day, month, year, doctor, hospital, patient, count, charge).

For all your calculations, show your work steps, i.e., if you just write down the results, you will not get any marks!

**Question 5** (4/20)

Assume that a data warehouse contains 20 dimensions, each with five levels of granularity.

- (a) Users are mainly interested in four particular dimensions, each having three frequently accessed levels for rolling up and drilling down. Design a data cube structure to efficiently support this preference, and justify your design.
- (b) A user wants to drill through the data cube, down to the raw data for one or two particular dimensions. How would you support this feature?