UNIVERSITY OF MALAYA

Midterm Test FOR THE DEGREE OF Master of Data Science

ACADEMIC SESSION 2021/2022 : SEMESTER 2

WQD 7005 : Data Mining

Time: 3 hours

INSTRUCTIONS TO CANDIDATES:

Answer **ALL** questions (100 marks).

1. Give your **own definition** of data warehousing. (10 marks)

(10 marks)

2. List 5 of your own data mining queries vs database processing queries.

3. You may use a group assignment data set as a base to create a data warehouse. Your design of the data warehouse must have at least **4** dimensions with at least **two** measures. If your current dataset does not meet the above requirements. You may recommend adding more attributes/columns into your current dataset.

   Draw a schema diagram for the above data warehouse using *Star* schema. Justify your attributes which you selected for each dimension.

   Hints: **irrelevant attributes** in the *S*tar schema, awarded mark will be deducted.

(30 marks)

4. You may use your **group assignment data sets t**o perform the followingactivities.

a) Use smoothing by bin boundaries to smooth the above data, you may determine **effective bin depth**. Illustrate your steps. Comment on the effect of this technique for the given data.

(10 marks)

(4 marks)

b) How might you determine outliers in the data?

c) What other methods are there for data smoothing?

(6 marks)

5. Redundant data occur often when integration of multiple databases, Redundant attributes may be able to be detected by correlation analysis. Kindly use group assignment data sets to perform the χ 2 statistic tests the hypothesis that A and B are independent. (example, A and B are from your own dataset.)

(Hints: Data transformation may be required in this process)

(30 marks)

END