COMP 5130/6130 Data Mining Fall 2021

Final 12/07/2021 12pm-2:30pm

Name:_____

1. (25 points) Suppose we have 9 records consist of the *age* and *hourly wage* data.

age	20	25	27	34	37	48	51	55	65
hourly wage(USD)	22	31	27	45	50	40	48	56	68

a. Partition the *age* data into 3 bins by **equal width** partitioning.

b. Calculate the **standard deviation** (round to the nearest integer) of *hourly wage* and use z-score normalization to transform the *hourly wage* 45.

c. Use normalization by decimal scaling to transform the *hourly wage* 45.

2. (25 points) Given a data warehouse with four dimensions *date, spectator, location*, and *movie*, and the two measures are *count* and *charge*. Spectators may be kids, students, or adults, where charge is the fare that a particular spectator pays when watching a particular movie in a particular cinema on a given date.

a. Draw a complete star schema diagram for this data warehouse.

b. Starting with the **base cuboid** [*date, spectator, location, movie*], what specific OLAP operations should you perform in order to list the total charge paid by *student* spectators to watch the movie *"Tenet"* at *AMC theater* in *2021*?

c. Which dimension is not suitable for bitmap indexing and why? (one dimension is enough)

3. (25 points) Below is a table from an employee database with four attributes. Assuming *status* is the class label attribute.

department	*status*	*age*	*salary*
sales	senior	31…35	46K…50K
sales	junior	26…30	36K…40K
sales	junior	31…35	36K…40K
systems	junior	21…25	46K…50K
systems	senior	31…35	66K…70K
systems	junior	26…30	46K…50K
systems	senior	41…45	66K…70K
marketing	senior	36…40	46K…50K
marketing	junior	31…35	41K…45K
secretary	senior	46…50	36K…40K
secretary	junior	26…30	36K…40K

a. Calculate the information gained by branching on attribute *salary.*

4. (25 points) A dataset has transactions as below. Let *minimum support* = 60%.

TID	*Items purchased*
T1	{yogurt, egg, kiwifruit, orange, cake}
T2	{apple, egg, milk, kiwifruit}
T3	{orange, kiwifruit, yogurt, egg, donut, noodle}
T4	{cake, milk, yogurt, kiwifruit}
T5	{orange, milk, noodle, kiwifruit, yogurt, egg}

a. Find all (not just one) frequent item sets using Apriori algorithm.