1. True/false questions analogous in style to those in the first problem of the first exam.

2. Consider a database with fact table Sales(model, color, date, dealer, cnt), similar to our in-class example. Attribute cnt is the measure attribute, and is intended to be the total number of automobiles sold for the given model, color, date, and dealer. Consider the expanded table CUBE(Sales), with the complete “data cube” for Sales.

(a) List what exact tuples of the cube we would use to answer the following queries. You will be deducted points for any unnecessary tuples that you list.

i. For each dealer, report the most popular car model for the dealer (i.e., the model for which the largest number of automobiles have been sold by the dealer).

Answer: All tuples of CUBE(Sales) that match pattern (m, *, *, d, c), where m, d, and c are non-* values.

ii. . . . (additional queries)

(b) Consider the following instance of a table Sales2, where cnt is the measure attribute:

<table>
<thead>
<tr>
<th>model</th>
<th>color</th>
<th>dealer</th>
<th>cnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Golf</td>
<td>red</td>
<td>D1</td>
<td>10</td>
</tr>
<tr>
<td>Jetta</td>
<td>red</td>
<td>D1</td>
<td>5</td>
</tr>
<tr>
<td>Jetta</td>
<td>silver</td>
<td>D2</td>
<td>100</td>
</tr>
<tr>
<td>Jetta</td>
<td>silver</td>
<td>D3</td>
<td>50</td>
</tr>
<tr>
<td>Golf</td>
<td>red</td>
<td>*</td>
<td>10</td>
</tr>
<tr>
<td>Jetta</td>
<td>red</td>
<td>*</td>
<td>5</td>
</tr>
<tr>
<td>Jetta</td>
<td>silver</td>
<td>*</td>
<td>150</td>
</tr>
<tr>
<td>Golf</td>
<td>*</td>
<td>D1</td>
<td>10</td>
</tr>
<tr>
<td>Jetta</td>
<td>*</td>
<td>D1</td>
<td>5</td>
</tr>
<tr>
<td>Jetta</td>
<td>*</td>
<td>D2</td>
<td>100</td>
</tr>
<tr>
<td>Jetta</td>
<td>*</td>
<td>D3</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>red</td>
<td>D1</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>silver</td>
<td>D2</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>silver</td>
<td>D3</td>
<td>50</td>
</tr>
<tr>
<td>Golf</td>
<td>*</td>
<td>*</td>
<td>10</td>
</tr>
<tr>
<td>Jetta</td>
<td>*</td>
<td>*</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>D1</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>D2</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>D3</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>red</td>
<td>*</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>silver</td>
<td>*</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>*</td>
<td>165</td>
</tr>
</tbody>
</table>

List the complete table CUBE(Sales2) for the instance of Sales2 above. Do not include any tuples for which cnt=0.

Answer:
3. Consider the following 6 “market baskets,” with items A, B, C, D, and E:

- A, B, C, D
- A, B
- B, C, D
- B, D, E
- A, B, C
- A, B, C, E

(a) Compute all itemsets with support 50% or higher.
Answer: ∅, {A}, {B}, {C}, {D}, {A, B}, {A, C}, {B, C}, {B, D}, {A, B, C}

(b) Consider each (large) itemset in the answer to the previous point that has two or more items, and list all association rules with only one item on the right hand side that you can generate using all the items in the itemset (i.e., all rules that you report should (1) involve two or more items and (2) have only one item on the right hand side). Report the support and confidence associated with each of these rules.
Answer:

<table>
<thead>
<tr>
<th>Association Rule</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A → B</td>
<td>2/3</td>
<td>1</td>
</tr>
<tr>
<td>B → A</td>
<td>2/3</td>
<td>2/3</td>
</tr>
<tr>
<td>A → C</td>
<td>1/2</td>
<td>3/4</td>
</tr>
<tr>
<td>C → A</td>
<td>1/2</td>
<td>3/4</td>
</tr>
<tr>
<td>B → C</td>
<td>2/3</td>
<td>2/3</td>
</tr>
<tr>
<td>C → B</td>
<td>2/3</td>
<td>1</td>
</tr>
<tr>
<td>B → D</td>
<td>1/2</td>
<td>1/2</td>
</tr>
<tr>
<td>D → B</td>
<td>1/2</td>
<td>1</td>
</tr>
<tr>
<td>A, B → C</td>
<td>1/2</td>
<td>3/4</td>
</tr>
<tr>
<td>A, C → B</td>
<td>1/2</td>
<td>1</td>
</tr>
<tr>
<td>B, C → A</td>
<td>1/2</td>
<td>3/4</td>
</tr>
</tbody>
</table>

4. A time-series clustering problem along the lines of the examples that Ioannis described during his lecture.

5. Consider a relation \( R(X, Y, Z) \) such that \( X \) forms the primary key of the relation, and a three-dimensional grid file on attributes \( X, Y, \) and \( Z \). Suppose that (1) the grid file was built by always partitioning the grid on the \( X \) and \( Y \) attributes, and never on the \( Z \) attribute; and (2) the grid directory is stored in main memory. Using the version of grid files that we discussed in class, indicate how many disk blocks we will need to read to answer each of the following queries; justify your answers carefully:

(a) SELECT *
FROM R
WHERE R.X=3 AND R.Y=5 AND R.Z=3
Answer: 1 block, because there is exactly one directory cell (with one associated disk block) that corresponds to \( X = 3 \) AND \( Y = 5 \) (AND \( Z=3 \)). (This is a point query.)

(b) ... (additional queries)