5.9 Show how to express $\text{group by cube}(a, b, c, d)$ using $\text{rollup}$; your answer should have only one $\text{group by}$ clause.

Answer:

$$\text{groupby rollup}(a), \text{rollup}(b), \text{rollup}(c), \text{rollup}(d)$$

5.11 Consider the $\text{sales}$ relation from Section 5.6. Write an SQL query to compute the cube operation on the relation, giving the relation in Figure 5.21. Do not use the cube construct.

Answer:

```
(select color, size, sum(number)
   from sales
   groupby color, size
 )
union
(select color, 'all', sum(number)
   from sales
   groupby color
 )
union
(select 'all', size, sum(number)
   from sales
   groupby size
 )
union
(select 'all', size, sum(number)
   from sales
   groupby size
 )
union
(select 'all', 'all', sum(number )
   from sales
 )
```
20.3 Suppose that there are two classification rules, one that says that people with salaries between $10,000 and $20,000 have a credit rating of good, and another that says that people with salaries between $20,000 and $30,000 have a credit rating of good. Under what conditions can the rules be replaced, without any loss of information, by a single rule that says people with salaries between $10,000 and $30,000 have a credit rating of good?

**Answer:** Consider the following pair of rules and their confidence levels:

<table>
<thead>
<tr>
<th>No.</th>
<th>Rule</th>
<th>Conf.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>$\forall$ persons $P. \ 10000 &lt; P.salary \leq 20000 \Rightarrow P.credit = \text{good}$</td>
<td>60%</td>
</tr>
<tr>
<td>2.</td>
<td>$\forall$ persons $P. \ 20000 &lt; P.salary \leq 30000 \Rightarrow P.credit = \text{good}$</td>
<td>90%</td>
</tr>
</tbody>
</table>

The new rule has to be assigned a confidence-level which is between the confidence-levels for rules 1 and 2. Replacing the original rules by the new rule will result in a loss of confidence-level information for classifying persons, since we cannot distinguish the confidence levels of people earning between 10000 and 20000 from those of people earning between 20000 and 30000. Therefore we can combine the two rules without loss of information only if their confidences are the same.

20.8 Construct a decision-tree classifier with binary splits at each node, using tuples in relation $r(A, B, C)$ shown below as training data; attribute C denotes the class. Show the final tree, and with each node show the best split for each attribute along with its information gain value.
\[(1, 2, a), (2, 1, a), (2, 5, b), (3, 3, b), (3, 6, b),
(4, 5, b), (5, 5, c), (6, 3, b), (6, 7, c)\]

**Answer:** Figure 20.1 shows one possible decision tree for the data. Using the Gini purity metric, the purity of the initial data set is

\[1 - \sum_{i=1}^{k} p_i^2 = 1 - \left(\frac{2}{9}\right)^2 + \left(\frac{5}{9}\right)^2 + \left(\frac{2}{9}\right)^2 = 0.595259\]

The first branch splits on \(B \leq 2\), giving a purity score of \(1 - 1^2 = 0\) for those attributes with \(B \leq 2\) (all are classified as \(a\)), and a purity score of

\[1 - \left(\frac{2}{7}\right)^2 + \left(\frac{5}{7}\right)^2 = 0.40816\]

for the remaining items. The weighted purity of the entire set is

\[\frac{2}{9} \times 0 + \frac{7}{9} \times 0.40816 = 0.31746\]

The information gain from this split is \(0.595259 - 0.31746 = 0.27513\). Next, we split on \(A < 5\). The 4 data items with \(A < 5\) all have class \(b\), and thus have purity 0. The remaining 3 items have purity

\[1 - \left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 = 0.44444\]

The weighted purity of these sets is

\[\frac{4}{7} \times 0 + \frac{3}{7} \times 0.44444 = 0.19048\]

The information gain from the second split is \(0.40816 - 0.19048 = 0.21769\).

Finally, we split on \(B \leq 3\). One data item satisfies this predicate and has class \(b\). The other two items both have class \(c\). The purity of these two sets is 0. The information gain from the final split is \(0.21769 - 0 = 0.21769\).
20.9 Suppose half of all the transactions in a clothes shop purchase jeans, and one third of all transactions in the shop purchase T-shirts. Suppose also that half of the transactions that purchase jeans also purchase T-shirts. Write down all the (nontrivial) association rules you can deduce from the above information, giving support and confidence of each rule.

\textbf{Answer}: The rules are as follows. The last rule can be deduced from the previous ones.

\begin{tabular}{|l|c|c|}
\hline
Rule & Support & Conf. \\
\hline
\forall \text{transactions } T, \text{true } \Rightarrow \text{buys}(T, \text{jeans}) & 50\% & 50\% \\
\forall \text{transactions } T, \text{true } \Rightarrow \text{buys}(T, \text{t-shirts}) & 33\% & 33\% \\
\forall \text{transactions } T, \text{buys}(T, \text{jeans}) \Rightarrow \text{buys}(T, \text{t-shirts}) & 25\% & 50\% \\
\forall \text{transactions } T, \text{buys}(T, \text{t-shirts}) \Rightarrow \text{buys}(T, \text{jeans}) & 25\% & 75\% \\
\hline
\end{tabular}