Assignment 9

Qinglin Li, 5110309074

Problem 1

a. **Hash-partitioning:** Too many records with the same value for the hashing attribute, or a poorly chosen hash function without the properties of randomness and uniformity, can result in a skewed partition. To improve the situation, we should experiment with better hashing functions for that relation.

b. **Range-partitioning:** Non-uniform distribution of values for the partitioning attribute (including duplicate values for the partitioning attribute) which are not taken into account by a bad partitioning vector is the main reason for skewed partitions. Sorting the relation on the partitioning attribute and then dividing it into n ranges with equal number of tuples per range will give a good partitioning vector with very low skew.

Problem 2

The 5 partitions are 1-20, 21-30, 31-50, 51-75 and 76-100.
Frequencies:

1. 1-20:$15 + 5 = 20$
2. 21-30:20
3. 31-50:$10 + 10 = 20$
4. 51-75:$5 + 5 + 20/10 * 5 = 20$
5. 76-100:$20/10 * 5 + 5 + 5 = 20$

Problem 3

groupby rollup(a), rollup(b), rollup(c), rollup(d)

Problem 4

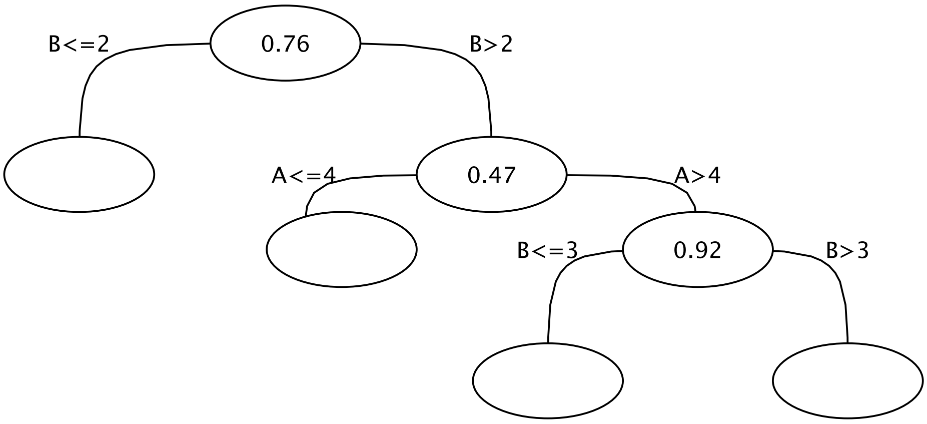
```
SELECT t, sum(c)
FROM (SELECT c, ntile(20) OVER (ORDER BY (a)) AS t FROM r) tt
GROUPBY t
```

Problem 5

```
SELECT 1, COUNT(*)
FROM account
WHERE 3 * balance <= (SELECT MAX (balance) FROM account)
UNION
SELECT 2, COUNT (*)
FROM account
WHERE 3 * balance > (SELECT MAX(balance) FROM account)
  AND 1.5 * balance <= (SELECT MAX(balance) FROM account)
UNION
SELECT 3, COUNT (*)
FROM account
WHERE 1.5 * balance > (SELECT MAX(balance) FROM account)
```

Problem 6

The information gain is the value in the node


B<=2
0.76
B>2
A<=4
0.47
A>4
B<=3
0.92
B>3