Question 1

(1):

Location	Time	Item	Quantity
Sydney	2005	PS2	1400
Sydney	2006	PS2	1500
Sydney	2006	Wii	500
Melbourne	2005	Xbox 360	1700
Sydney	2005	ALL	1400
Sydney	2006	ALL	2000
Melbourne	2005	ALL	1700
Sydney	ALL	PS2	2900
Sydney	ALL	Wii	500
Melbourne	ALL	Xbox 360	1700
ALL	2005	PS2	1400
ALL	2006	PS2	1500
ALL	2006	Wii	500
ALL	2005	Xbox 360	1700
Sydney	ALL	ALL	3400
Melbourne	ALL	ALL	1700
ALL	2005	ALL	3100
ALL	2006	ALL	2000
ALL	ALL	PS2	2900
ALL	ALL	Wii	500
ALL	ALL	Xbox 360	1700
ALL	ALL	ALL	5100

(2):

```sql
SELECT  Location, Time, Item, SUM(Quantity)
FROM    Sales
GROUP BY    Location, Time, Item
UNION ALL
SELECT  Location, Time, ALL, SUM(Quantity)
FROM    Sales
GROUP BY    Location, Time
UNION ALL
SELECT  Location, ALL, Item, SUM(Quantity)
FROM    Sales
GROUP BY    Location, Item
UNION ALL
SELECT  ALL, Time, Item, SUM(Quantity)
FROM    Sales
GROUP BY    Time, Item
```

```sql
UNION ALL
SELECT  Location, ALL, ALL, SUM(Quantity)
FROM    Sales
GROUP BY    Location
UNION ALL
SELECT  ALL, Time, ALL, SUM(Quantity)
FROM    Sales
GROUP BY    Time
UNION ALL
SELECT  ALL, ALL, Item, SUM(Quantity)
FROM    Sales
GROUP BY    Item
UNION ALL
SELECT  ALL, ALL, ALL, SUM(Quantity)
FROM    Sales
```

(3):

Location	Time	Item	Quantity
Sydney	2006	ALL	2000
Sydney	ALL	PS2	2900
ALL	ALL	PS2	2900
ALL	2005	ALL	3100
ALL	2006	ALL	2000
Sydney	ALL	ALL	3400
ALL	ALL	ALL	5100

(4):

The mapping function: $f_{Location,Time,Item}(x) = 12 * f_{Location}(x) + 4 * f_{Time}(x) + f_{Item}(x)$

Location	Time	Item	$offset$ ($f_{Location,Time,Item}(x)$)	Dense MD array (Quantity)
1	1	1	17	1400
1	2	1	21	1500
1	2	3	23	500
2	1	2	30	1700
1	1	0	16	1400
1	2	0	20	2000
2	1	0	28	1700
1	0	1	13	2900
1	0	3	15	500
2	0	2	26	1700
0	1	1	5	1400
0	2	1	9	1500
0	2	3	11	500

0	1	2	6	1700
1	0	0	12	3400
2	0	0	24	1700
0	1	0	4	3100
0	2	0	8	2000
0	0	1	1	2900
0	0	3	3	500
0	0	2	2	1700
0	0	0	0	5100

Question 2

	p_1	p_2	p_3	p_4	p_5
p_1	1.00	0.10	0.41	0.55	0.35
p_2		1.00	0.64	0.47	0.98
p_3			1.00	0.44	0.85
p_4				1.00	0.76
p_5					1.00

$Step1: Merge\ the\ two\ closest\ clusters. The\ max\ similarity\ is\ similarity(cluster_2, cluster_5) =$ 0.98

$Step2: Update\ the\ similarity\ matrix\ by\ group\ average.$

$$similarity(cluster_{2,5}, cluster_1) = \frac{\sum_{p_i,p_j \in p_1,p_2,p_5} similarity(p_i,p_j)}{(|cluster_{2,5}| + |cluster_1|) * (|cluster_{2,5}| + |cluster_1| - 1)}$$
$$= 2 * \frac{0.1 + 0.35 + 0.98}{3 * 2} \approx 0.48$$

$$similarity(cluster_{2,5}, cluster_3) = \frac{\sum_{p_i,p_j \in p_3,p_2,p_5} similarity(p_i,p_j)}{(|cluster_{2,5}| + |cluster_3|) * (|cluster_{2,5}| + |cluster_3| - 1)}$$
$$= 2 * \frac{0.64 + 0.98 + 0.85}{3 * 2} \approx 0.82$$

$$similarity(cluster_{2,5}, cluster_4) = \frac{\sum_{p_i,p_j \in p_4,p_2,p_5} similarity(p_i,p_j)}{(|cluster_{2,5}| + |cluster_4|) * (|cluster_{2,5}| + |cluster_4| - 1)}$$
$$= 2 * \frac{0.64 + 0.98 + 0.85}{3 * 2} \approx 0.77$$

	p_1	p_3	p_4	$p_{2,5}$
p_1	1.00	0.41	0.55	0.48
p_3		1.00	0.44	0.85
p_4			1.00	0.77
p_5				1.00

$Step3: Merge\ the\ two\ closest\ clusters. The\ max\ similarity\ is\ similarity(cluster_{2,5}, cluster_3) =$ 0.85

*Step*4: *Update the similarity matrix by group average.*

$$similarity\left(cluster_{2,5,3}, cluster_1\right) = \frac{\sum_{p_i,p_j \in p_1,p_2,p_3,p_5} similarity(p_i,p_j)}{\left(\left|cluster_{2,5,3}\right| + |cluster_1|\right) * \left(\left|cluster_{2,5,3}\right| + |cluster_1| - 1\right)}$$

$$= 2 * \frac{0.1 + 0.41 + 0.35 + 0.64 + 0.98 + 0.85}{4 * 3} \approx 0.56$$

$$similarity\left(cluster_{2,5,3}, cluster_4\right) = \frac{\sum_{p_i,p_j \in p_4,p_2,p_3,p_5} similarity(p_i,p_j)}{\left(\left|cluster_{2,5,3}\right| + |cluster_4|\right) * \left(\left|cluster_{2,5,3}\right| + |cluster_4| - 1\right)}$$

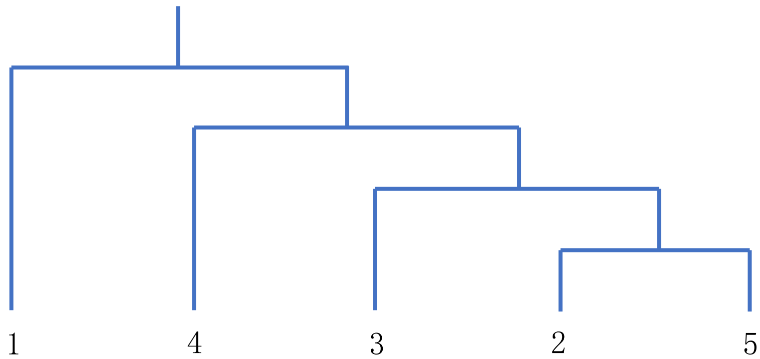
$$= 2 * \frac{0.64 + 0.47 + 0.98 + 0.44 + 0.85 + 0.76}{4 * 3} \approx 0.69$$

	$\boldsymbol{p_1}$	$\boldsymbol{p_4}$	$\boldsymbol{p_{2,3,5}}$
p_1	1.00	0.55	0.56
p_4		1.00	0.69
p_5			1.00

*Step*5: *Merge the two closest clusters. The max similarity is* $similarity\left(cluster_{2,3,5}, cluster_4\right) =$ 0.69

*Step*6: *Merge the last two clusters,* $cluster_{2,3,4,5}$ *and* $cluster_1$.

The final results:



Question 3

(1):

$\boldsymbol{Data}$: D *is a dataset of* n $d-$*dimensional points*; k *is the number of clusters.*

Initialize k *centers* $C = [c_1, c_2, \ldots, c_k]$;
$canStop \leftarrow \boldsymbol{false}$;
$\boldsymbol{while}$ $canStop = \boldsymbol{false}$ $\boldsymbol{do}$
    Initialize k *empty clusters* $G = [g_1, g_2, \ldots, g_k]$;
    $\boldsymbol{for\ each}$ *data point* $p \in D$ $\boldsymbol{do}$
        $c_x \leftarrow NearestCenter(p, C)$;
        $g_{c_x}.append(p)$;
    $\boldsymbol{end\ for}$
    $canStop \leftarrow \boldsymbol{true}$;
    $\boldsymbol{for\ each}$ *group* $g \in G$ $\boldsymbol{do}$
        $tempc = c_i$
        $c_i \leftarrow ComputeCenter(g)$
        $\boldsymbol{if}$ $tempc \neq c_i$ *then*
            $canStop \leftarrow \boldsymbol{false}$;
        $\boldsymbol{end\ if}$
    $\boldsymbol{end\ for}$
return G;

(2):

For each point in each iteration, there are three situation:

1. *If the point is still belong to the cluster and the centers point would not change in this iteration:* $cost(g_i)$ *will not change so the cost function will not increases.*
2. *If the point is still belong to the cluster and the centers point changes in this iteration:* *It means that the new center point get a smaller* $cost(g_i)$ *so the cost of* k *cluster will not increase*
3. *If the point is not still belong to the cluster anymore in this iteration:* *It means that the point is more close to another centerpoint, we can get that the increase of* $cost(g_j) <$ *the decraese of* $cost(g_i)$ (*the point move from* g_i *to* g_j). *So the* $cost(g_1, g_2, \ldots, g_k)$ *is decreased.*

Combine with these three situations, we can get that the cost of k *cluster will not increase.*

(3):

There can be a example for $k-$*means algorithm converges to a local minima:*

If we are trying to find 2 *appropriate clustares for* $A = \{1,2,3,4,5\}$, *if we set* $c_1 = \{1,2\}$ *and* $c_2 = \{3,4,5\}$, *we will get the same objective value as* $c_1 = \{1,2,3\}$ *and* $c_2 = \{4,5\}$.