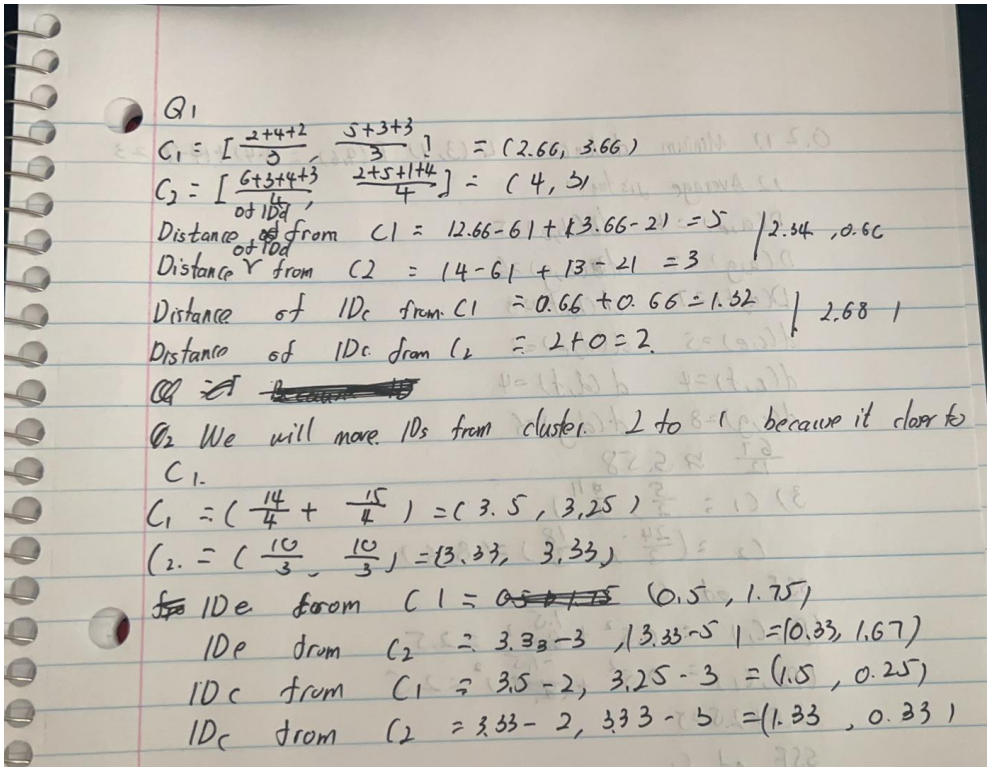
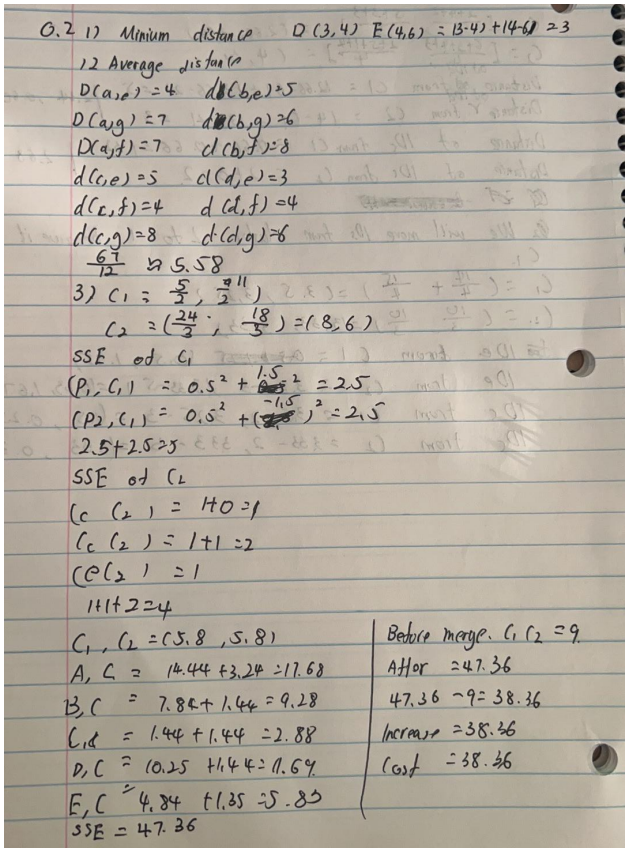
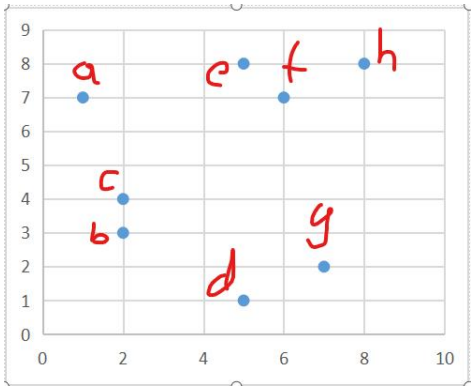
Q1

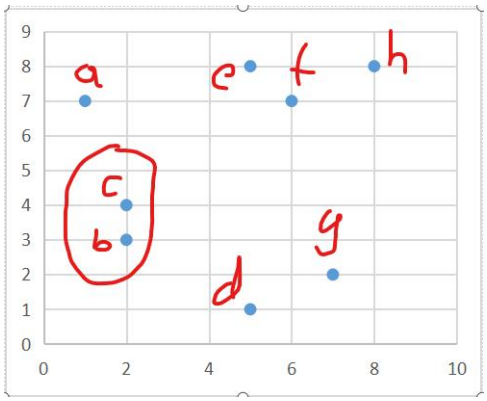

Q1
C1 = [(2+4+2)/3, (5+3+3)/3] = (2.66, 3.66)
C2 = [(6+3+4+3)/4, (2+5+1+4)/4] = (4, 3)
Distance of ID2 from C1 = 12.66-6| + |3.66-2| = 5
Distance of ID2 from C2 = |4-6| + |3-2| = 3
Distance of ID6 from C1 = 0.66 + 0.66 = 1.32
Distance of ID6 from C2 = 2+0 = 2.
Q2 We will move IDs from cluster 2 to 1 because it closer to C1.
C1 = (14/4 + 15/4) = (3.5, 3.25)
C2 = (10/3, 10/3) = (3.33, 3.33)
ID2 from C1 = (0.5, 1.75)
ID2 from C2 = |3.33-3|, |3.33-5| = (0.33, 1.67)
ID6 from C1 = 3.5-2, 3.25-3 = (1.5, 0.25)
ID6 from C2 = 3.33-2, 3.33-3 = (1.33, 0.33)

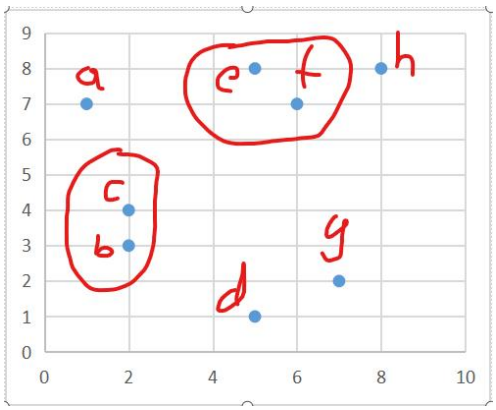
Q2

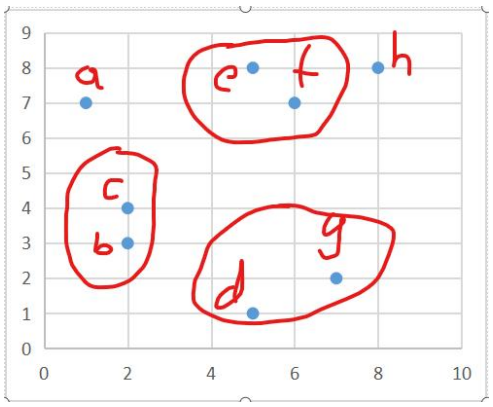

Q.2 1) Minium distance D(3,4) E(4,6) = |3-4| + |4-6| = 3
2) Average distance
D(a,e) = 4 d(b,e) = 5
D(a,g) = 7 d(b,g) = 6
D(a,f) = 7 d(b,f) = 8
d(c,e) = 5 d(d,e) = 3
d(c,f) = 4 d(d,f) = 4
d(c,g) = 8 d(d,g) = 6
67/12 ≈ 5.58
3) C1 = (5/2, 9/2)
C2 = (24/3, 18/3) = (8, 6)
SSE of C1
(P1, C1) = 0.5² + 1.5² = 2.5
(P2, C1) = 0.5² + (-1.5)² = 2.5
2.5+2.5=5
SSE of C2
(c C2) = 1+0 = 1
(c C2) = 1+1 = 2
(c C2) = 1
1+1+2=4
C1, C2 = (5.8, 5.8)
A, C = 14.44 + 3.24 = 17.68
B, C = 7.84 + 1.44 = 9.28
C, d = 1.44 + 1.44 = 2.88
D, C = 10.25 + 1.44 = 11.69
E, C = 4.84 + 1.35 = 5.89
SSE = 47.36
Before merge, C1 C2 = 9
After = 47.36
47.36 - 9 = 38.36
Increase = 38.36
Cost = 38.36

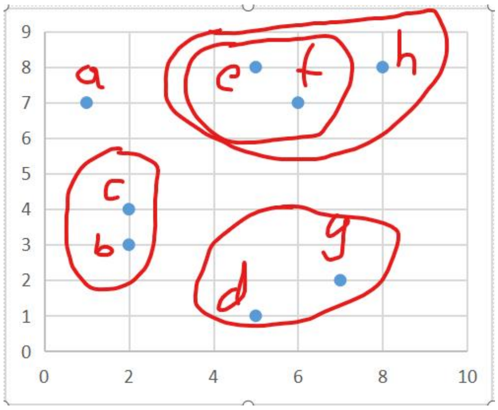
Q3

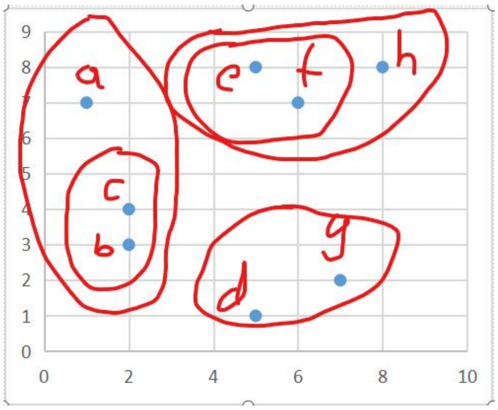

a
e
f
h
c
b
g
d
0 1 2 3 4 5 6 7 8 9
0 2 4 6 8 10

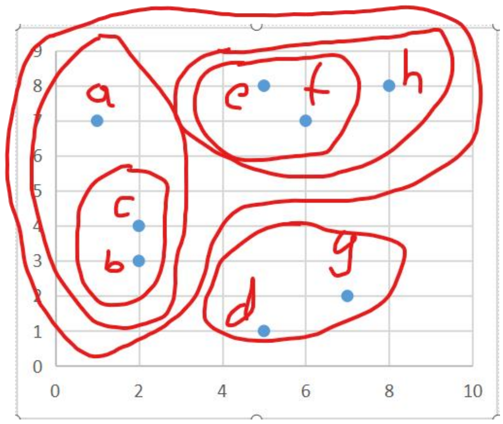

a
e
f
h
c
b
g
d
0 1 2 3 4 5 6 7 8 9
0 2 4 6 8 10

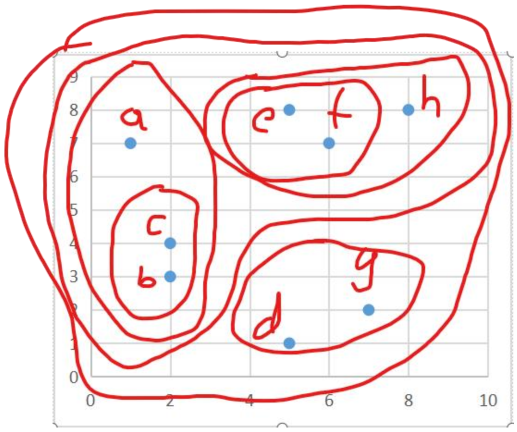

a
e
f
h
c
b
g
d
0 1 2 3 4 5 6 7 8 9
0 2 4 6 8 10


a
e
f
h
c
b
g
d
0 1 2 3 4 5 6 7 8 9
0 2 4 6 8 10









(a)(b)(c)(d)(e)(f)(g)(h)

(a)(b,c)(d)(e)(f)(g)(h)

(a)(b,c)(d)(e,f)(g)(h)

(a)(b,c)(d,g)(e,f)(h)

(a)(b,c)(d,g)(e,f,h)

(a,b,c)(d,g)(e,f,h)

Q4

(1)

K=2

```
Number of iterations: 6
Within cluster sum of squared errors: 33.30854661169795

Initial starting points (random):

Cluster 0: 61.5,391.34
Cluster 1: 92.1,393.25

Missing values globally replaced with mean/mode

Final cluster centroids:
                      Cluster#
Attribute    Full Data          0          1
               (506.0)    (202.0)    (304.0)
============================================
A1             68.5749    37.8658    88.9803
A2             356.674   389.2163   335.0506
```

K=3

```
kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 11.529148066305805

Initial starting points (random):

Cluster 0: 61.5,391.34
Cluster 1: 92.1,393.25
Cluster 2: 95.6,60.72

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1          2
               (506.0)    (191.0)    (275.0)     (40.0)
=============================================================
A1             68.5749    36.3461    87.7833      90.41
A2             356.674   388.9075    376.937    63.4515
```

K=4

```
Number of iterations: 17
Within cluster sum of squared errors: 6.895127380017731

Initial starting points (random):

Cluster 0: 61.5,391.34
Cluster 1: 92.1,393.25
Cluster 2: 95.6,60.72
Cluster 3: 95.6,396.9

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1          2          3
               (506.0)    (125.0)    (120.0)     (40.0)    (221.0)
=================================================================
A1             68.5749    27.3688     60.735      90.41    92.1864
A2             356.674   388.1318    390.291    63.4515   373.6995
```

K=5

```
Number of iterations: 13
Within cluster sum of squared errors: 5.355515572616415

Initial starting points (random):

Cluster 0: 61.5,391.34
Cluster 1: 92.1,393.25
Cluster 2: 95.6,60.72
Cluster 3: 95.6,396.9
Cluster 4: 73.3,385.91

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4
               (506.0)    (125.0)     (32.0)     (36.0)     (196.0)    (117.0)
=============================================================================
A1              68.5749    27.3688    94.0031    89.3444    91.8112    60.3274
A2             356.674    388.1318   278.3478    50.1542   385.6866   390.1994
```

K=6

```
kMeans
======

Number of iterations: 22
Within cluster sum of squared errors: 3.923262095570305

Initial starting points (random):

Cluster 0: 61.5,391.34
Cluster 1: 92.1,393.25
Cluster 2: 95.6,60.72
Cluster 3: 95.6,396.9
Cluster 4: 73.3,385.91
Cluster 5: 83.2,390.7

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4          5
               (506.0)     (91.0)     (28.0)     (36.0)     (165.0)     (93.0)     (93.0)
=================================================================================================
A1              68.5749    22.8352    94.0357    89.3444    94.303     47.6624    72.8914
A2             356.674    388.692   270.8661    50.1542   383.7804   389.0449   389.3692
```

K=7

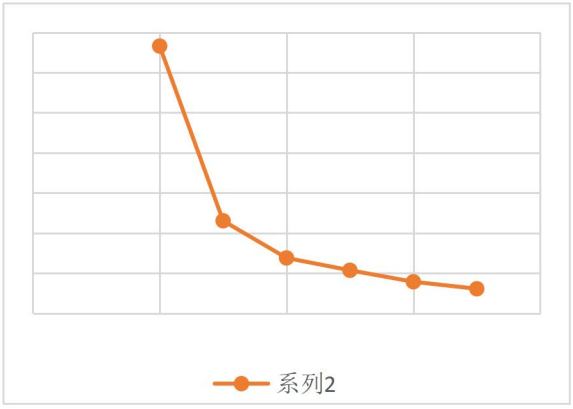
```
Number of iterations: 37
Within cluster sum of squared errors: 3.051532898594132

Initial starting points (random):

Cluster 0: 61.5,391.34
Cluster 1: 92.1,393.25
Cluster 2: 95.6,60.72
Cluster 3: 95.6,396.9
Cluster 4: 73.3,385.91
Cluster 5: 83.2,390.7
Cluster 6: 58,396.9

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4          5
               (506.0)     (84.0)     (28.0)     (36.0)     (154.0)     (72.0)     (82.0)
=================================================================================================
A1              68.5749    36.0667    94.0357    89.3444    94.9896    56.7236    76.9122
A2             356.674    389.7713   270.8661    50.1542   383.4029   390.3379   388.7613
```



optimal number of clusters:k=2

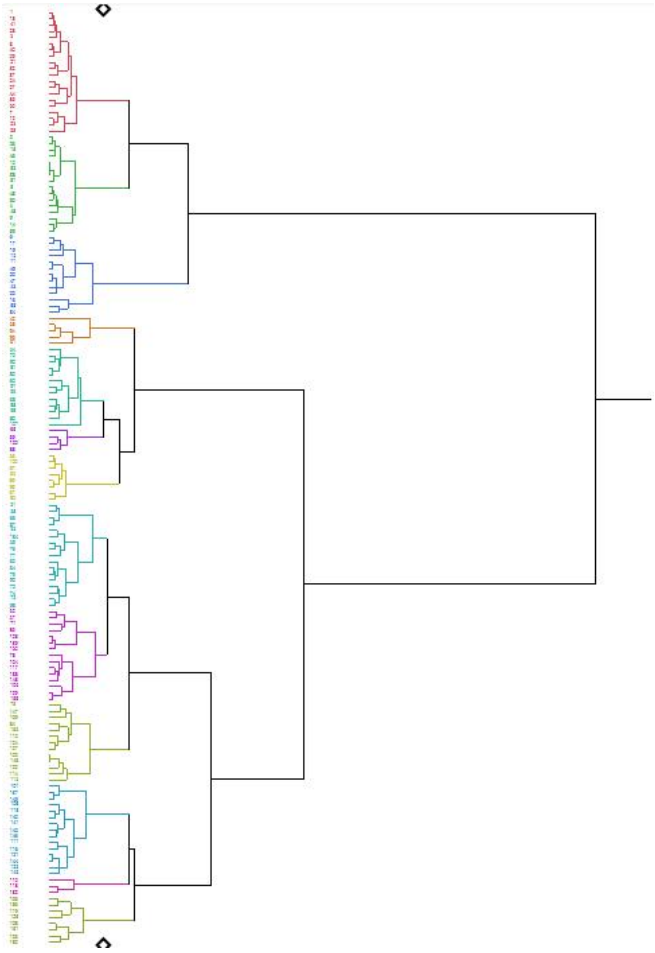
(2)

		1	2	3	4
Calories	Mean	156.5629	214.8264	171.026	167.5224
	STD	+/-11.1385	+/-15.5421	+/-18.191	+/-8.2984
	MAX	211.073064	196.0569717	211.073064	24.10206
	Min	135.9771213	135.9771213	166.0206963	135.9771213
Fiber	Mean	3.9537	7.2813	24.8333	0.225
	STD	+/-4.3567	+/-4.3536	+/-6.3456	+/-1.0665
	MAX	24.5	24.5	8.5	35.5
	Min	-0.5	-0.5	-0.5	-0.5
Sugars	Mean	10.9923	26.7789	26.0188	30.7703
	STD	+/-4.2619	+/-7.5774	+/-10.1438	+/-6.3388
	MAX	38.97474109	38.97474109	28.27701184	42.97815422
	Min	7.150685103	11.95246912	14.32746912	7.150685103
Potassium	Mean	81.2963	146.875	206.6667	46.75
	STD	+/-32.1233	+/-44.7962	+/-36.697	+/-24.4559
	MAX	230	190	100	260
	Min	20	30	35	35

So,cluster 1 has the healthiest cereals

Q5

(1)



10	2.43807889	54	56
9	2.77492066	1	2
8	2.78163606	51	73
7	2.79180266	101	110
6	2.98563726	101	106
5	2.99442891	42	54
4	4.87847370	1	6
3	5.64618841	51	101
2	8.91626260	42	51
1	19.16712684	1	42

I will chose No.4 because since No.4, the distance has a huge change

(2)

聚类比较

方法	聚类数	CCC	最佳
K 均值聚类	1	0	
K 均值聚类	2	3.31726	
K 均值聚类	3	4.93767	最优 CCC
K 均值聚类	4	3.57791	
K 均值聚类	5	2.10824	
K 均值聚类	6	1.77615	
K 均值聚类	7	2.69263	
K 均值聚类	8	2.96036	
K 均值聚类	9	3.83904	
K 均值聚类	10	4.49637	

K 均值聚类数=3

逐列统一尺度

聚类汇总

聚类	计数	步进	准则
1	55	8	0
2	46		
3	49		

聚类均值

聚类	sepallength	sepalwidth	petallength	petalwidth
1	6.69636364	3.06	5.41818182	1.93818182
2	5.70434783	2.63478261	4.21521739	1.3326087
3	5.01632653	3.44081633	1.46734694	0.24285714

聚类标准差

聚类	sepallength	sepalwidth	petallength	petalwidth
1	0.50342956	0.25768197	0.62960042	0.33493061
2	0.41648452	0.25978033	0.69625314	0.30861144
3	0.34484361	0.34518162	0.17188956	0.1069045