**Problem 1**

*Cluster 1 Centroid:*

ID: 1, 2, 3

*Cluster 2 Centroid:*

ID: 4, 5, 6, 7

*ID:1*

Distance to Cluster 1 = |3-4.67|+|4-3.67| = 1.67+0.33 = 2

Distance to Cluster 2 = |3-5.75|+|4-5.5| = 2.75+1.5 = 4.25

So, ID:1 is in Cluster 1.

*ID:2*

Distance to Cluster 1 = |5-4.67|+|3-3.67| = 0.33+0.67 = 1

Distance to Cluster 2 = |5-5.75|+|3-5.5| = 0.75+2.5 = 3.25

So, ID:2 is in Cluster 1.

*ID:3*

Distance to Cluster 1 = |6-4.67|+|4-3.67| = 1.33+0.33 = 1.66

Distance to Cluster 2 = |6-5.75|+|4-5.5| = 0.25+1.5 = 1.75

So, ID:3 is in Cluster 1.

*ID:4*

Distance to Cluster 1 = |4-4.67|+|5-3.67| = 0.67+1.33 = 2

Distance to Cluster 2 = |4-5.75|+|5-5.5| = 1.75+0.5 = 2.25

So, ID:4 is in Cluster 1.

*ID:5*

Distance to Cluster 1 = |4-4.67|+|7-3.67| = 0.67+3.33 = 4

Distance to Cluster 2 = |4-5.75|+|7-5.5| = 1.75+1.5 = 3.25

So, ID:5 is in Cluster 2.

*ID:6*

Distance to Cluster 1 = |7-4.67|+|6-3.67| = 2.33+2.33 = 4.66

Distance to Cluster 2 = |7-5.75|+|6-5.5| = 1.25+0.5 = 1.75

So, ID:6 is in Cluster 2.

*ID:7*

Distance to Cluster 1 = |8-4.67|+|4-3.67| = 3.33+0.33 = 3.66

Distance to Cluster 2 = |8-5.75|+|4-5.5| = 2.25+1.5 = 3.75

So, ID:7 is in Cluster 1.

Finally, we can get the clusters at the end of iterations:

Cluster1: Id:1,2,3,4,7

Cluster2: Id:5,6

**Problem 2**

Between a and e: |2-6|+|4-4| = 4

Between a and f: |2-7|+|4-2| = 7

Between a and g: |2-8|+|4-5| = 7

Between b and e: |3-6|+|6-4| = 5

Between b and f: |3-7|+|6-2| = 8

Between b and g: |3-8|+|6-5| = 8

Between c and e: |3-6|+|2-4| = 5

Between c and f: |3-7|+|2-2| = 6

Between c and g: |3-8|+|2-5| = 8

Between d and e: |4-6|+|3-4| = 3

Between d and f: |4-7|+|3-2| = 4

Between d and g: |4-8|+|3-5| = 6

***Maximum Distance:***we can find that the maximum distance between 2 clusters is equal to 8

***Mean Distance:***

So, the mean distance between 2 clusters is equal to 5.92, rounded to 2 decimal

**Problem 3**

***Distances:***

Dist(a, b) = |1-2| + |7-3| = 5

Dist(a, c) = |1-2| + |7-4| = 4

Dist(a, d) = |1-5| + |7-1| = 10

Dist(a, e) = |1-5| + |7-8| = 5

Dist(a, f) = |1-6| + |7-7| = 5

Dist(a, g) = |1-7| + |7-2| = 11

Dist(a, h) = |1-8| + |7-8| = 8

Dist(b, c) = |2-2| + |3-4| = 1

Dist(b, d) = |2-5| + |3-1| = 5

Dist(b, e) = |2-5| + |3-8| = 8

Dist(b, f) = |2-6| + |3-7| = 8

Dist(b, g) = |2-7| + |3-2| = 6

Dist(b, h) = |2-8| + |3-8| = 11

Dist(c, d) = |2-5| + |4-1| = 4

Dist(c, e) = |2-5| + |4-8| = 7

Dist(c, f) = |2-6| + |4-7| = 7

Dist(c, g) = |2-7| + |4-2| = 5

Dist(c, h) = |2-8| + |4-8| = 10

Dist(d, e) = |5-5| + |1-8| = 7

Dist(d, f) = |5-6| + |1-7| = 7

Dist(d, g) = |5-7| + |1-2| = 3

Dist(d, h) = |5-8| + |1-8| = 10

Dist(e, f) = |5-6| + |8-7| = 2

Dist(e, g) = |5-7| + |8-2| = 4

Dist(e, h) = |5-8| + |8-8| = 3

Dist(f, g) = |6-7| + |7-2| = 4

Dist(f, h) = |6-8| + |7-8| = 3

Dist(g, h) = |7-8| + |2-8| = 7

Initially, each point is in its own cluster:

Cluster 1: {a}

Cluster 2: {b}

Cluster 3: {c}

Cluster 4: {d}

Cluster 5: {e}

Cluster 6: {f}

Cluster 7: {g}

Cluster 8: {h}

***1st Merge***

The smallest distance is between b and c,which is equal to 1, so we merge them into a single cluster:

Cluster 1: {a}

Cluster 2: {b, c}

Cluster 3: {d}

Cluster 4: {e}

Cluster 5: {f}

Cluster 6: {g}

Cluster 7: {h}

***2nd Merge***

Next smallest distance is between e and f,which is equal to 2, merge them:

Cluster 1: {a}

Cluster 2: {b, c}

Cluster 3: {d}

Cluster 4: {e, f}

Cluster 5: {g}

Cluster 6: {h}

***3rd Merge***

The next smallest distance is between d and g,which is equal to 3, merge them:

Cluster 1: {a}

Cluster 2: {b, c}

Cluster 3: {d, g}

Cluster 4: {e, f}

Cluster 5: {h}

***4th Merge***

Next, we merge a with the cluster {b, c} because the minimum distance between a and {b, c} is 4 (a and c):

Cluster 1: {a, b, c}

Cluster 2: {d, g}

Cluster 3: {e, f}

Cluster 4: {h}

***5th Merge***

Merge cluster {e, f} with {h} because the minimum distance between these two clusters is 3 (f and h):

Cluster 1: {a, b, c}

Cluster 2: {d, g}

Cluster 3: {e, f, h}

***6th Merge***

Merge cluster {a, b, c} with {d, g} because the minimum distance between these two clusters is 4 (c and d):

Cluster 1: {a, b, c, d, g}

Cluster 2: {e, f, h}

***7th Merge(Final Merge)***

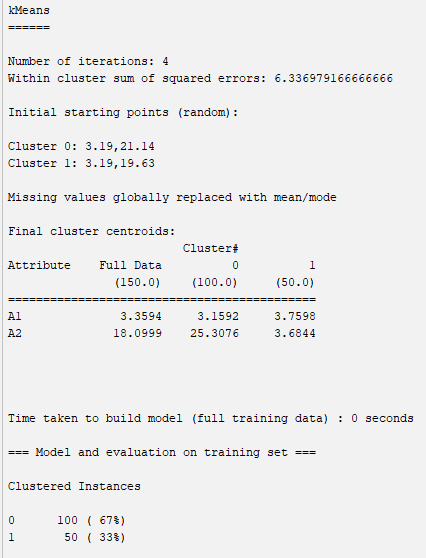
we merge the remaining clusters to get a single cluster:

Cluster 1: {a, b, c, d, e, f, g, h}

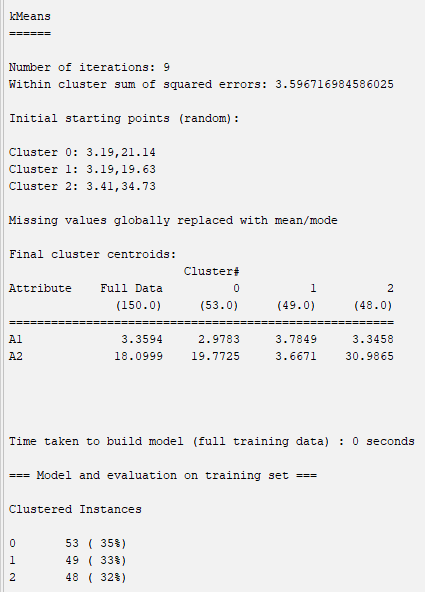
**Problem 4**

1. **graphed generated by Excel p4-1.xlsx, other from Weka**

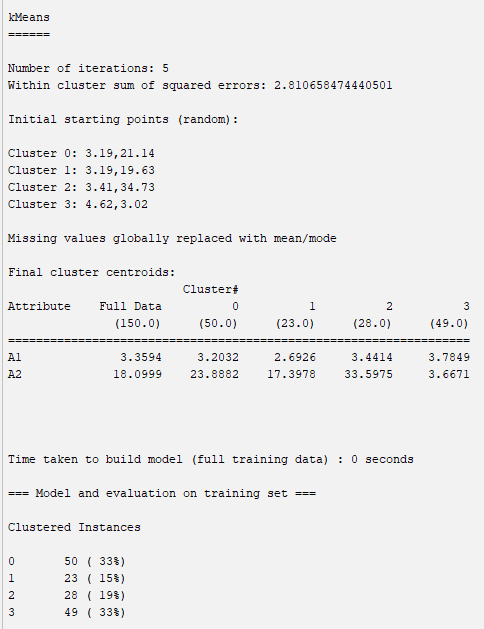
K=2



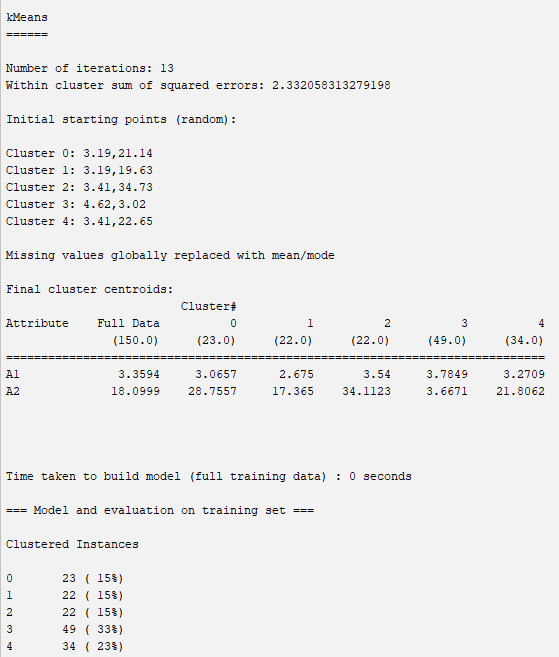
K=3



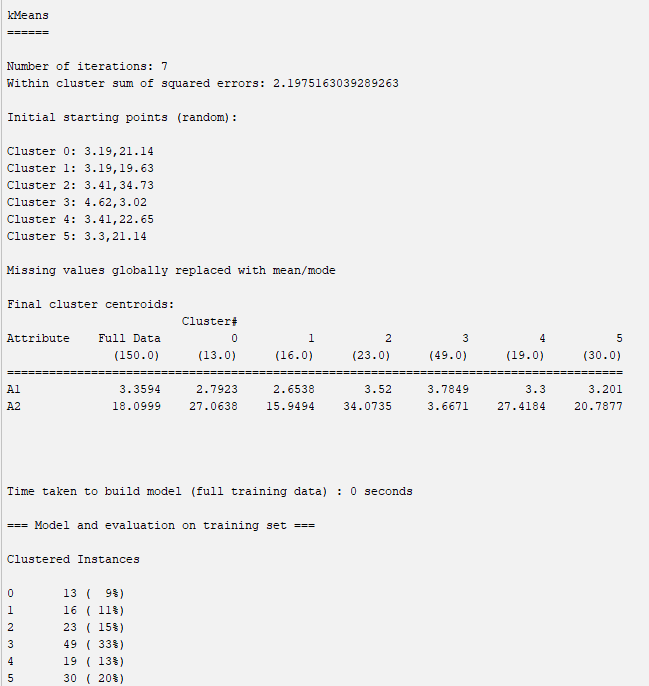
K=4



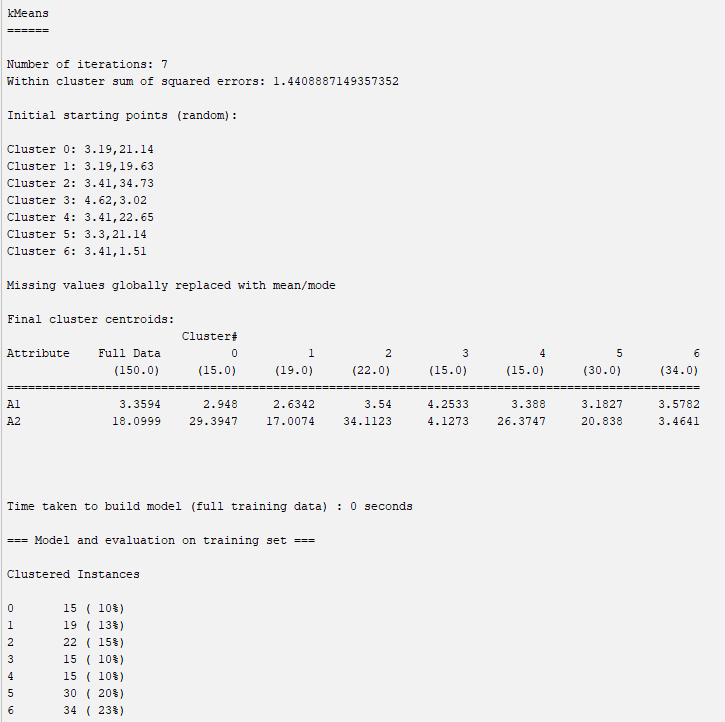
K=5



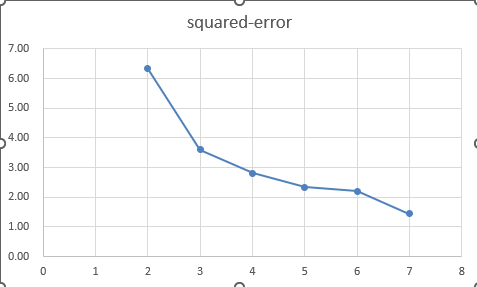
K=6



K=7



|  |  |
| --- | --- |
| k | squared-error |
| 2 | 6.34 |
| 3 | 3.60 |
| 4 | 2.81 |
| 5 | 2.33 |
| 6 | 2.20 |
| 7 | 1.44 |



By this graph and Elbow Method, we can find that the optimal number of clusters is 3

**(2)**

**A:**

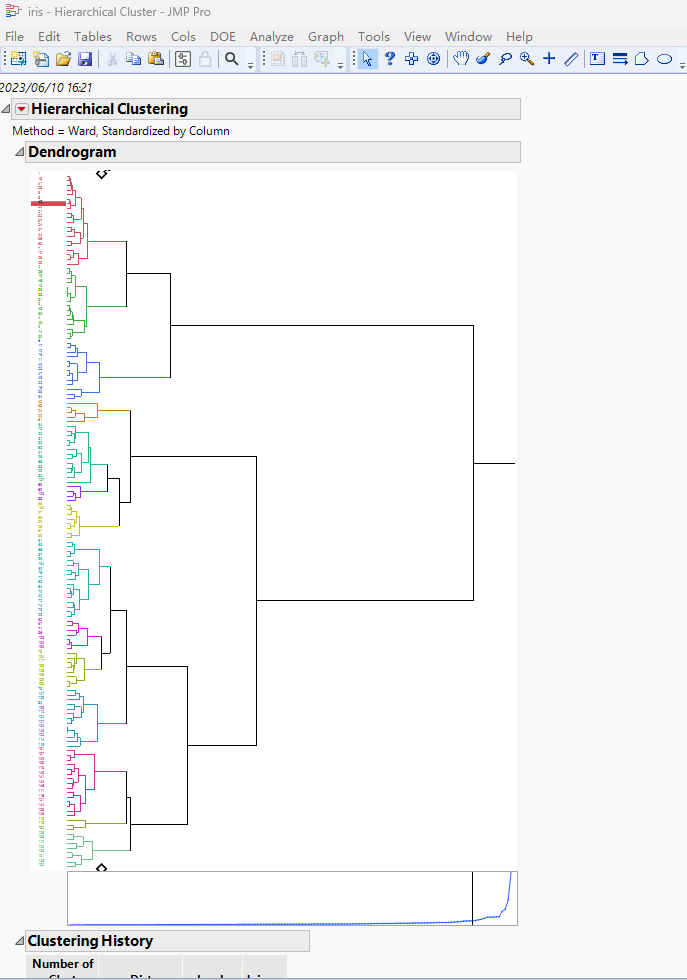
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Clusters | | | |
|  | | 1 | 2 | 3 | 4 |
| Calories | mean | 98.64865 | 63.333333 | 110.9091 | 130 |
|  | stddev | +/-14.5606 | +/-11.547 | +/-5.2636 | +/-16.4751 |
|  | max | 120 | 70 | 120 | 160 |
|  | min | 50 | 50 | 100 | 100 |
| Fiber | mean | 1.856757 | 11 | 0.590909 | 3.4 |
|  | stddev | +/-1.3716 | +/-2.6458 | +/-0.648 | +/-1.3654 |
|  | max | 5 | 14 | 2 | 6 |
|  | min | 0 | 9 | 0 | 1.5 |
| Sugars | mean | 3.243243 | 3.6666667 | 11.31818 | 10.266667 |
|  | stddev | +/-2.1939 | +/-3.2146 | +/-2.056 | +/-2.4044 |
|  | max | 7 | 6 | 15 | 14 |
|  | min | 0 | 0 | 8 | 7 |
| Potassium | mean | 80.40541 | 310 | 45.22727 | 166.66667 |
|  | stddev | +/-37.8866 | +/-26.4575 | +/-25.8582 | +/-49.7733 |
|  | max | 190 | 330 | 120 | 260 |
|  | min | 15 | 280 | 25 | 95 |

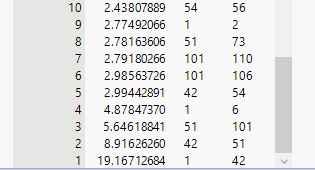
**B:**

So,Cluster 2 has the healthiest cereals.

**Problem 5**

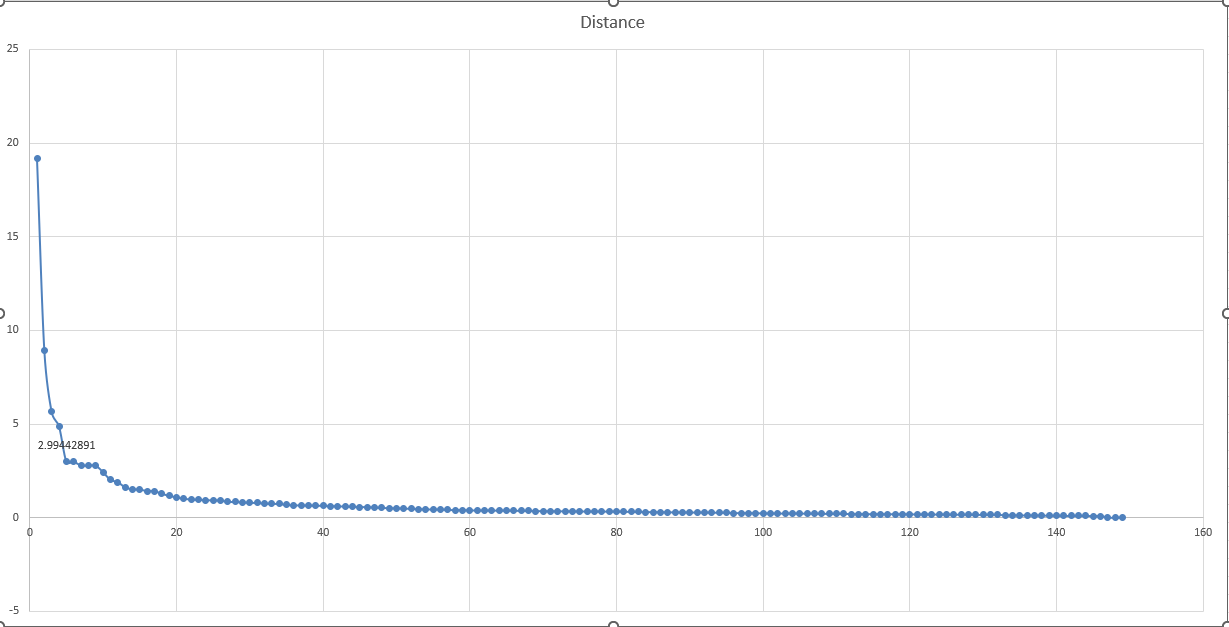
Hierarchical Clustering



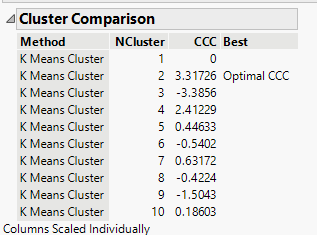
****

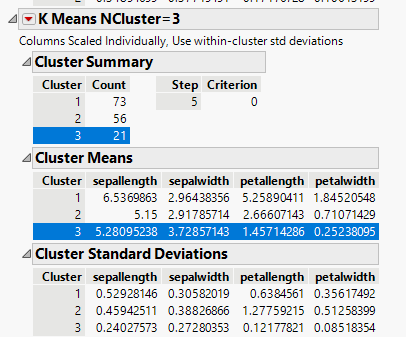
***10: Excel: p5.xlsx***

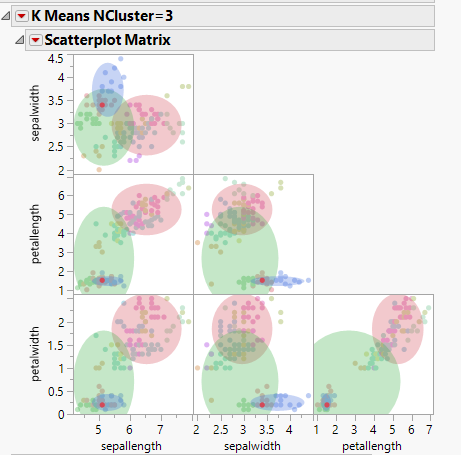
***I listed all the clustering history into excel, and generated the graph, according to Elbow Method, the optimal number of clustering is 5.***

******

K Means

******

******

******