Yiduo Feng

CS 699

Assignment 6

Date: 6/21/2022

**Problem 1 (20 points)**. The k-means algorithm, which we discussed in the class, is being run on a small two-dimensional dataset. After a certain number of iterations, you have two clusters as shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| ID | x | y | Cluster |
| 1 | 3 | 4 | Cluster 1 |
| 2 | 5 | 3 | Cluster 1 |
| 3 | 6 | 4 | Cluster 1 |
| 4 | 4 | 5 | Cluster 2 |
| 5 | 4 | 7 | Cluster 2 |
| 6 | 7 | 6 | Cluster 2 |
| 7 | 8 | 4 | Cluster 2 |

Run one more iteration of the k-Means clustering algorithm and show the two clusters at the end of the iteration. Use Manhattan distance when calculating the distance between objects.

C1.x = (3+5+6)/3 = 4.666667

C1.y = (4+3+4)/3 = 3.666667

C2.x = (4+4+7+8)/4 = 5.75

C2.y = (5+7+6+4)/4 = 5.5

Manhattan distance of each point:

1: |3-4.666667|+|4-3.666667| = 2 Cluster 1

|3-5.57|+|4-5.5| = 4.25 Cluster 2

2: |5 -4.666667|+|3-3.666667|= 1 Cluster 1

|5 -5.57|+|3-5.5|= 3.25 Cluster 2

3: |6 -4.666667|+|4 -3.666667|=1.6667 Cluster 1

|6 -5.57|+|4-5.5|= 1.75 Cluster 2

4: |4-4.666667|+| 5-3.666667| =2 Cluster 1

|4-5.75|+| 5-5.5 | = 2.25 Cluster 2

5: |4-4.666667|+| 7-3.666667| =4 Cluster 1

|4-5.75|+| 7-5.5 | = 3.25 Cluster 2

6: |7-4.666667|+| 6-3.666667| = 4. 6667 Cluster 1

|7-5.75|+| 6-5.5 | = 1.75 Cluster 2

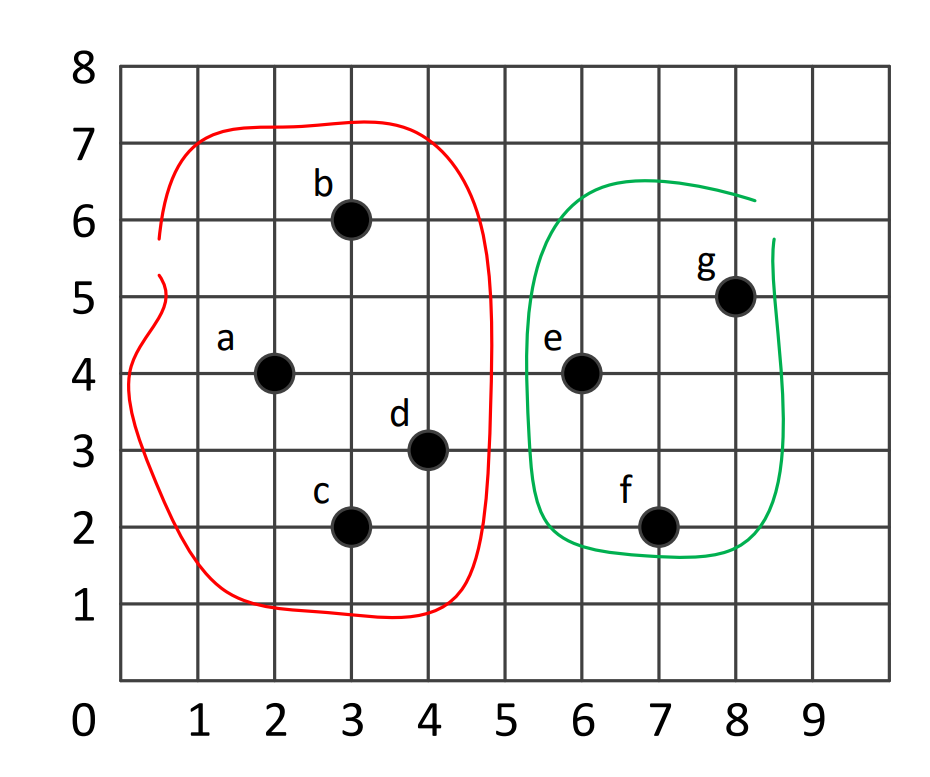
7: |8-4.666667|+| 4-3.666667| = 3.6667 Cluster 1

|8-5.75|+| 4-5.5 | = 3.75 Cluster2

Thus, the new cluster is:

|  |  |  |  |
| --- | --- | --- | --- |
| ID | x | y | Cluster |
| 1 | 3 | 4 | Cluster 1 |
| 2 | 5 | 3 | Cluster 1 |
| 3 | 6 | 4 | Cluster 1 |
| 4 | 4 | 5 | Cluster 1 |
| 5 | 4 | 7 | Cluster 2 |
| 6 | 7 | 6 | Cluster 2 |
| 7 | 8 | 4 | Cluster 1 |

**Problem 2 (20 points).** Consider the following two clusters:



Compute the distance between the two clusters (1) using the maximum distance method and (2) using the mean distance method. Use the Manhattan distance measure when calculating the distance between objects.

1. maximum distance

d(a,e) = |2-8|+|4-5| =7

d(a,f) = |2-7|+|4-2|=7

d(a,g) = |2-6|+|4-4|=4

d(b,e) = |3-8|+|6-5|=6

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

d(b,g) = |3-6|+|6-4|=5

d(c,e) = |3-8|+|2-5|=8

d(c,f) = |3-7|+|2-2|=4

d(c,g) = |3-6|+|2-4|=5

d(d,e) = |4-8|+|3-5|=6

d(d,f) = |4-7|+|3-2|=4

d(d,g) = |4-6|+|3-4|=3

so the maximum distance is 8.

1. mean distance

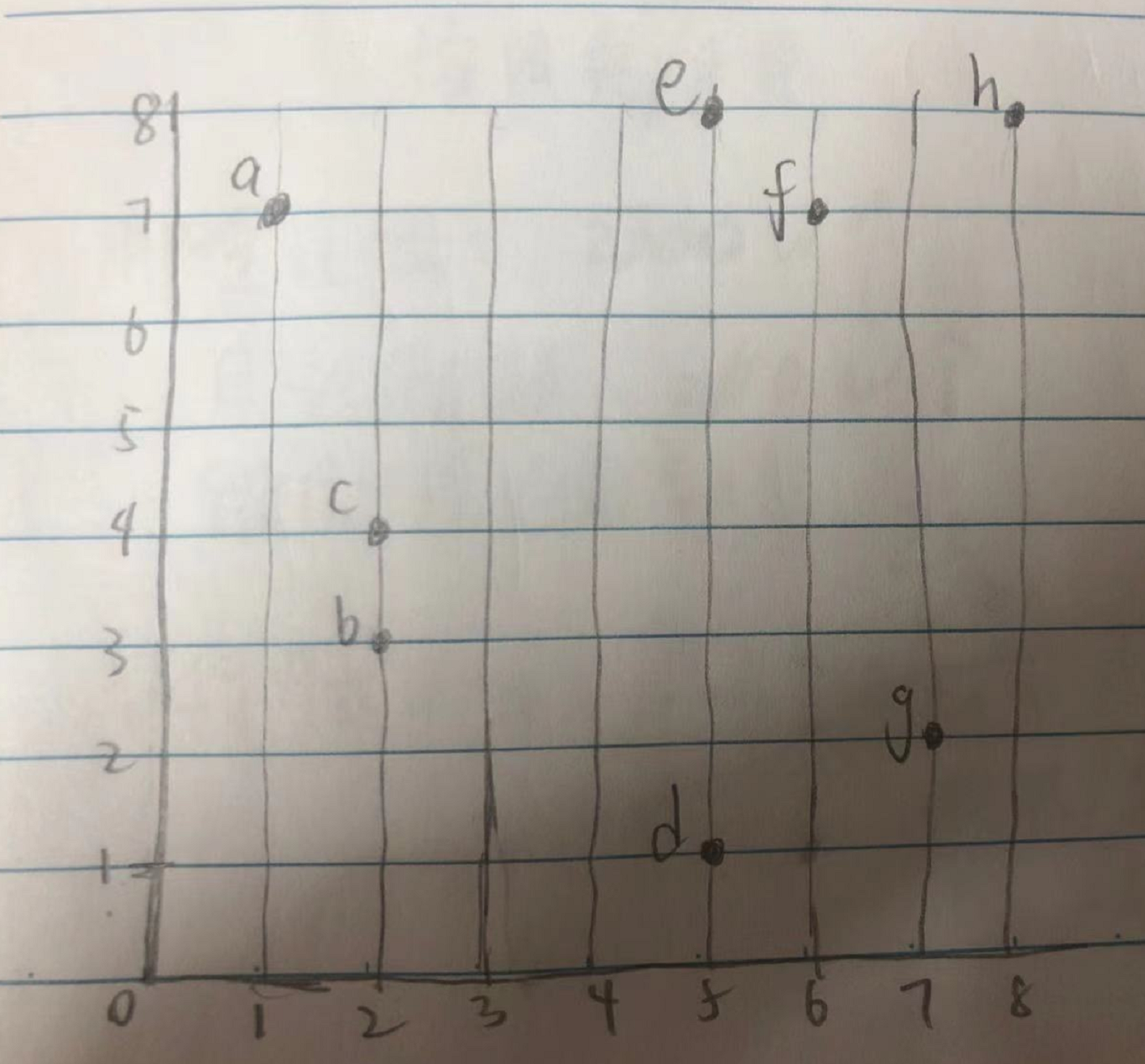
According to the calculation in (1)

Average distance = (7+7+4+6+8+5+8+4+5+6+4+3)/12 = 5.583333

**Problem 3 (20 points)**. Consider the following dataset, which has eight 2-dimensional objects.

|  |  |  |
| --- | --- | --- |
| Object | x | y |
| a | 1 | 7 |
| b | 2 | 3 |
| c | 2 | 4 |
| d | 5 | 1 |
| e | 5 | 8 |
| f | 6 | 7 |
| g | 7 | 2 |
| h | 8 | 8 |

Using the agglomerative hierarchical clustering approach that we discussed in the class, create two clusters. Use the *minimum distance* method with the Manhattan distance measure. You need to show, at each step, which two clusters are merged. You must also show, at each step, all clusters and their objects at the step. You must decide which two clusters are merged yourself and you must not use any software to do that.



Step 1:

According to the graph, b and c are nearest.

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

7 clusters

{a}{b,c}{d}{e}{f}{g}{h}

Step 2:

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

6 clusters

{a}{b,c}{d}{e, f}{g}{h}

Step 3:

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

5 clusters

{a}{b,c}{d, g}{e, f}{h}

Step 4:

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

4 clusters

{a}{b, c}{d, g}{e, f, h}

Step 5:

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

3 clusters

{a, b, c}{d, g}{e, f, h}

Step 6:

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

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

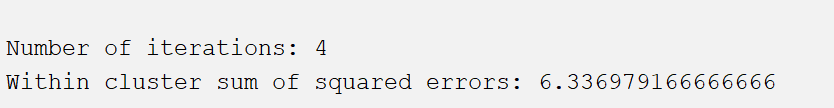
2 clusters

{a, b, c, d, g}{e, f, h}

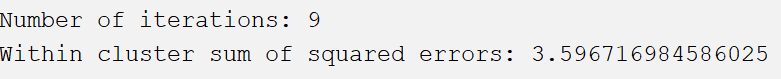
**Problem 4** **(20 points)**.

(1). Use the *a6-p4-1.csv* dataset for this problem. It has 2 attributes and 150 tuples. Run the *SimpleKMeans* algorithm of Weka on this dataset with *k* = 2, 3, 4, 5, 6, and 7. For each *k*, record the value of *within cluster sum of squared errors* (which you can find in Weka's cluster output window) and plot a graph where the x-axis is *k* and y-axis is *within cluster sum of squared errors*. Then, determine an optimal number of clusters using the *elbow method* which we discussed in the class.

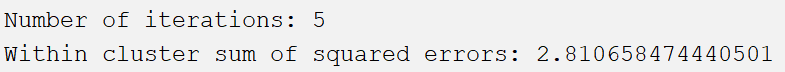
K=2:



K=3:



K=4:



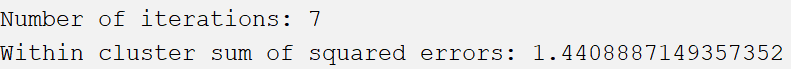
K=5:



K=6:



K=7:



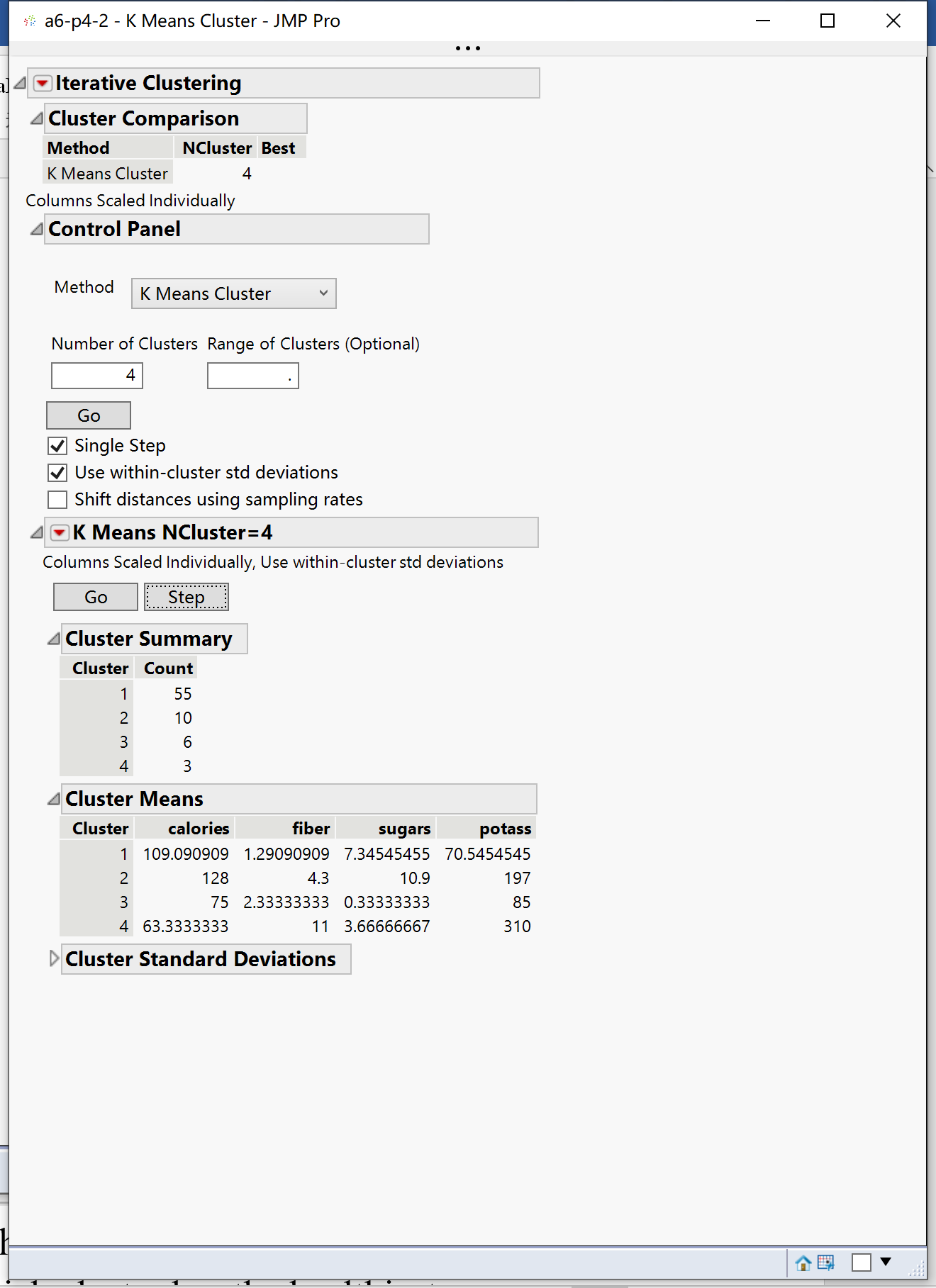
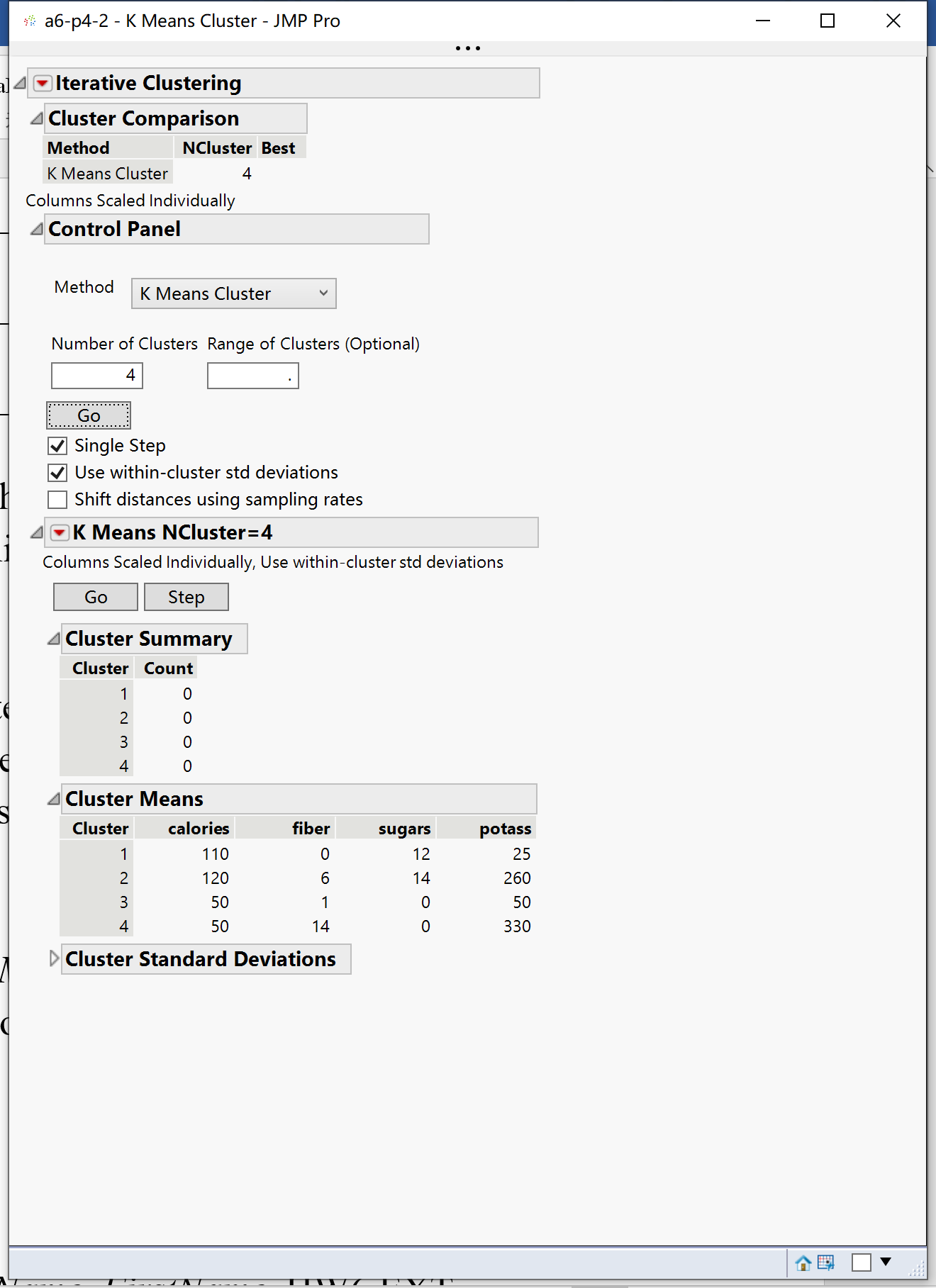
Thus, According to the graph, 3 is optimal number of clusters by using the *elbow method.*

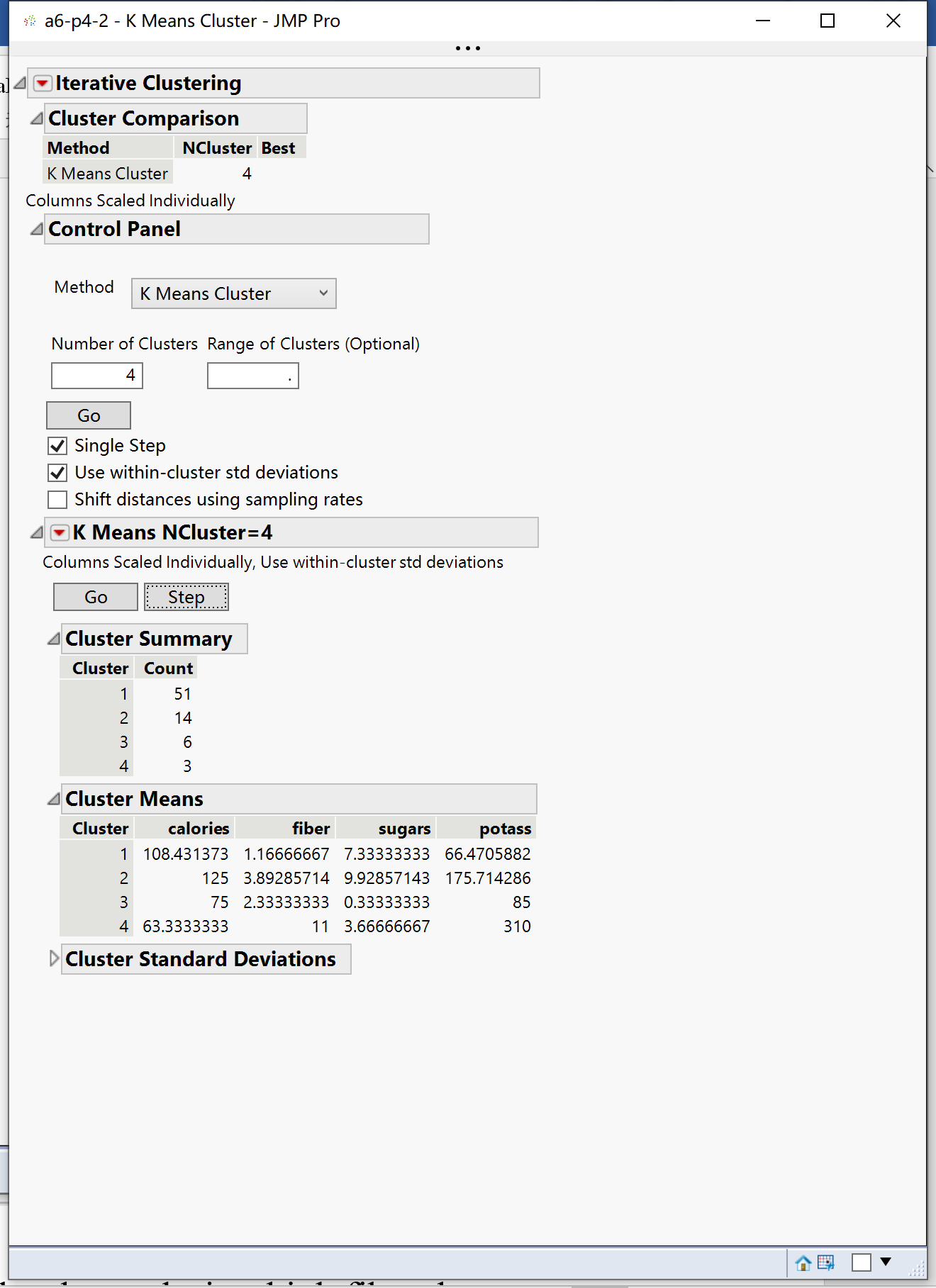
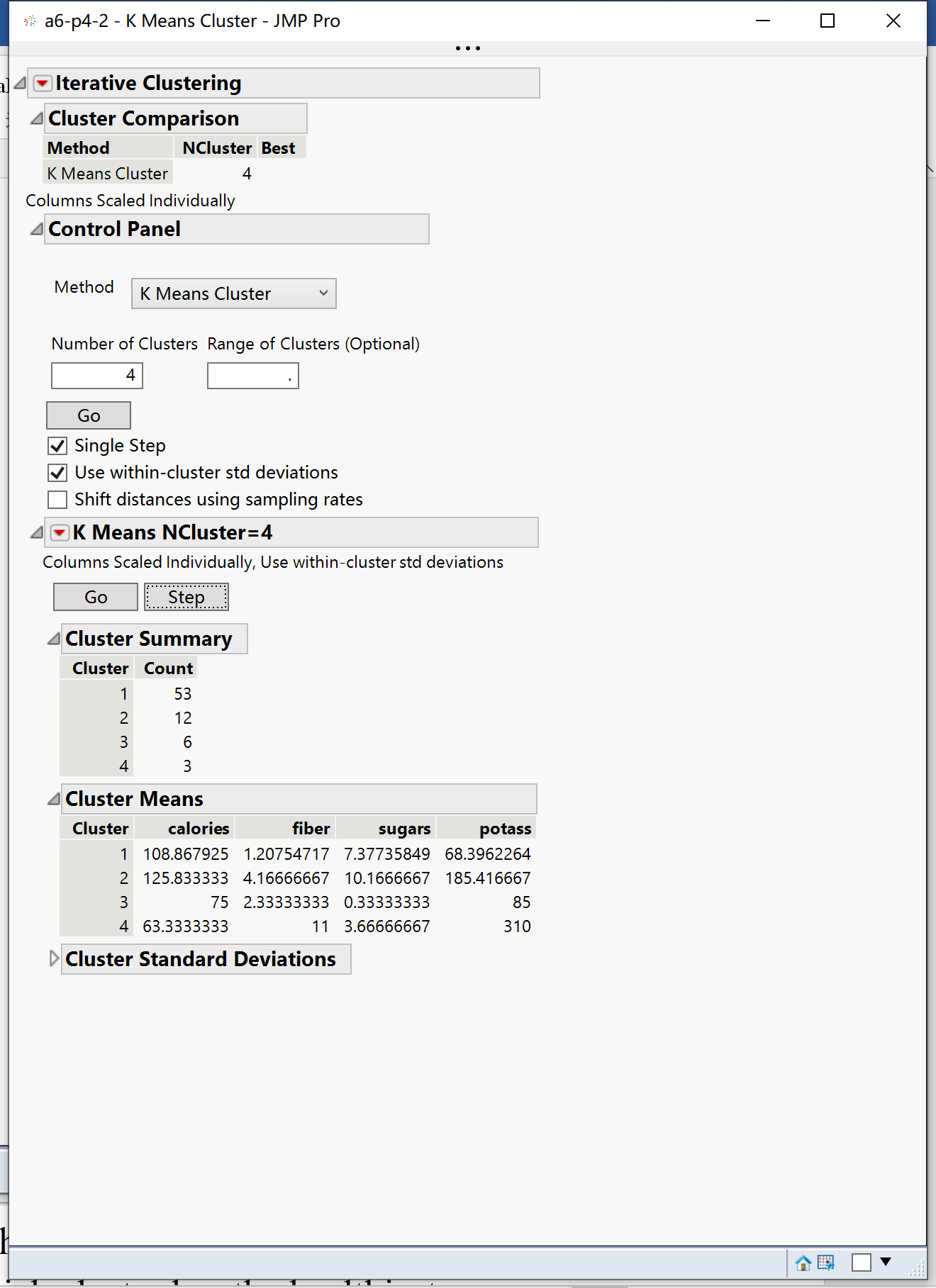
(2). Use the *a6-p4-2.csv* dataset for this problem, which has 4 attributes and 77 tuples. This dataset was downloaded from JMP sample dataset library and modified for this assignment. The dataset has nutritional information for 77 breakfast cereals. Each tuple in the dataset represents a cereal product and the four attributes are *calories*, *fiber*, *sugars*, and *potassium*.

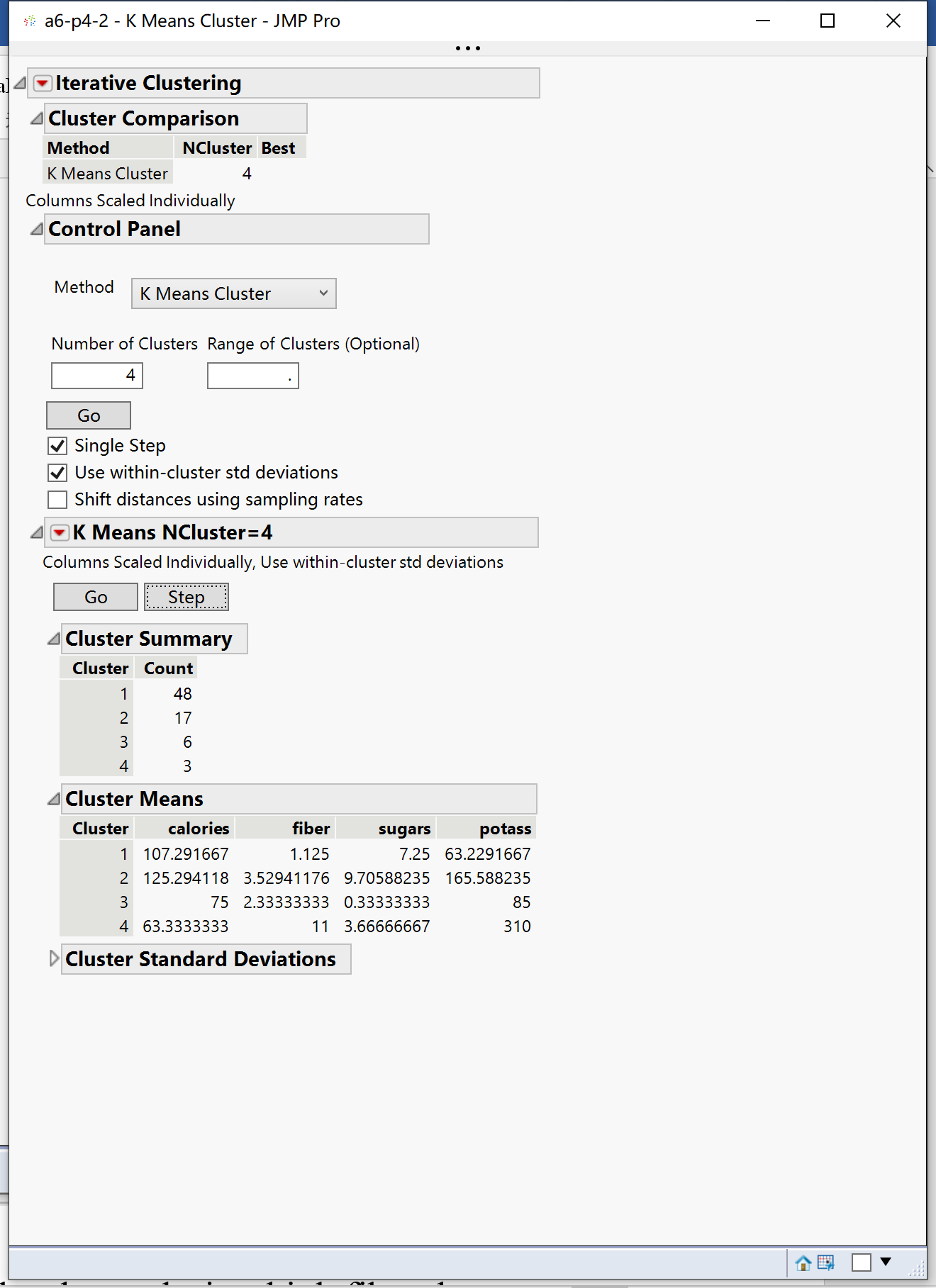
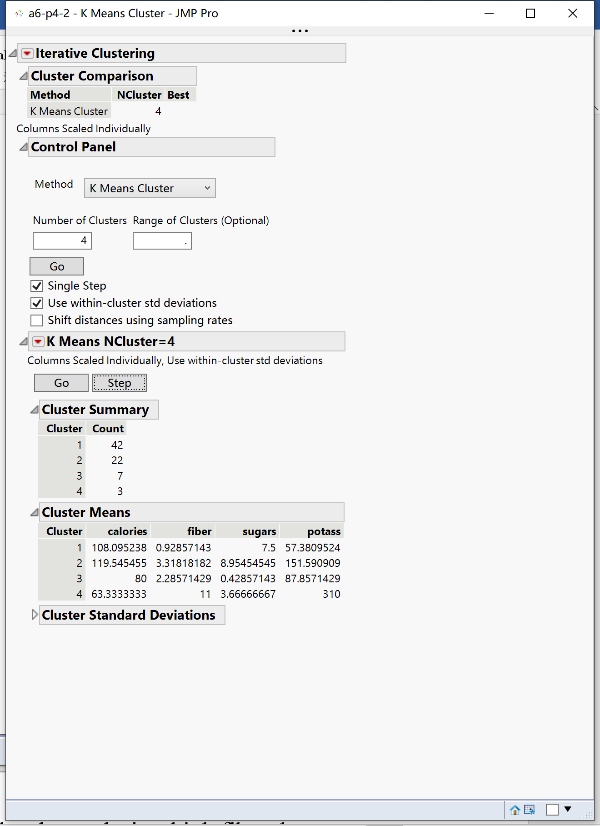
(2)-a. Run the *SimpleKMeans* algorithm of Weka on this dataset with *k* = 4 and prepare the following table that summarizes the profiles of four clusters:

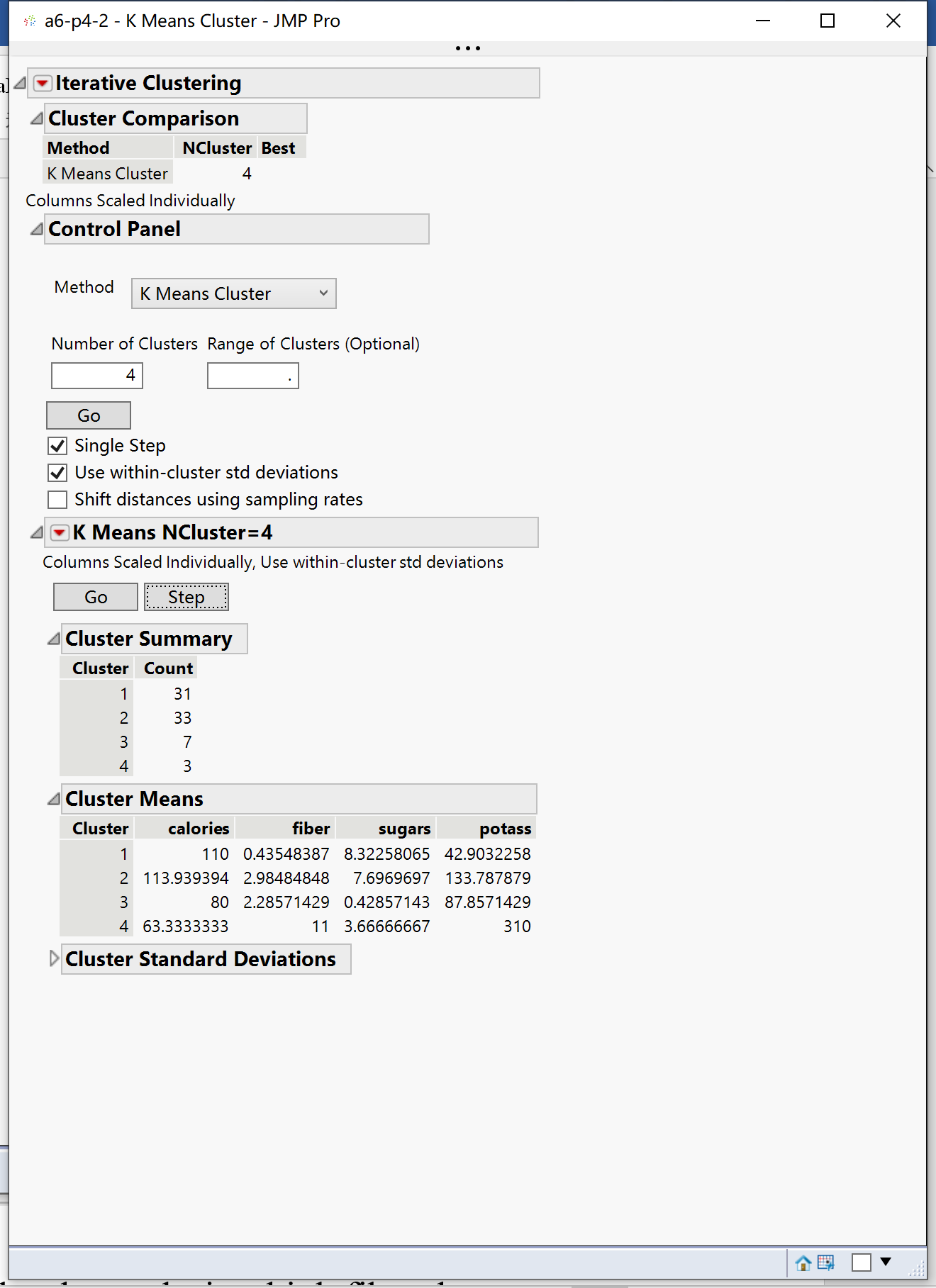
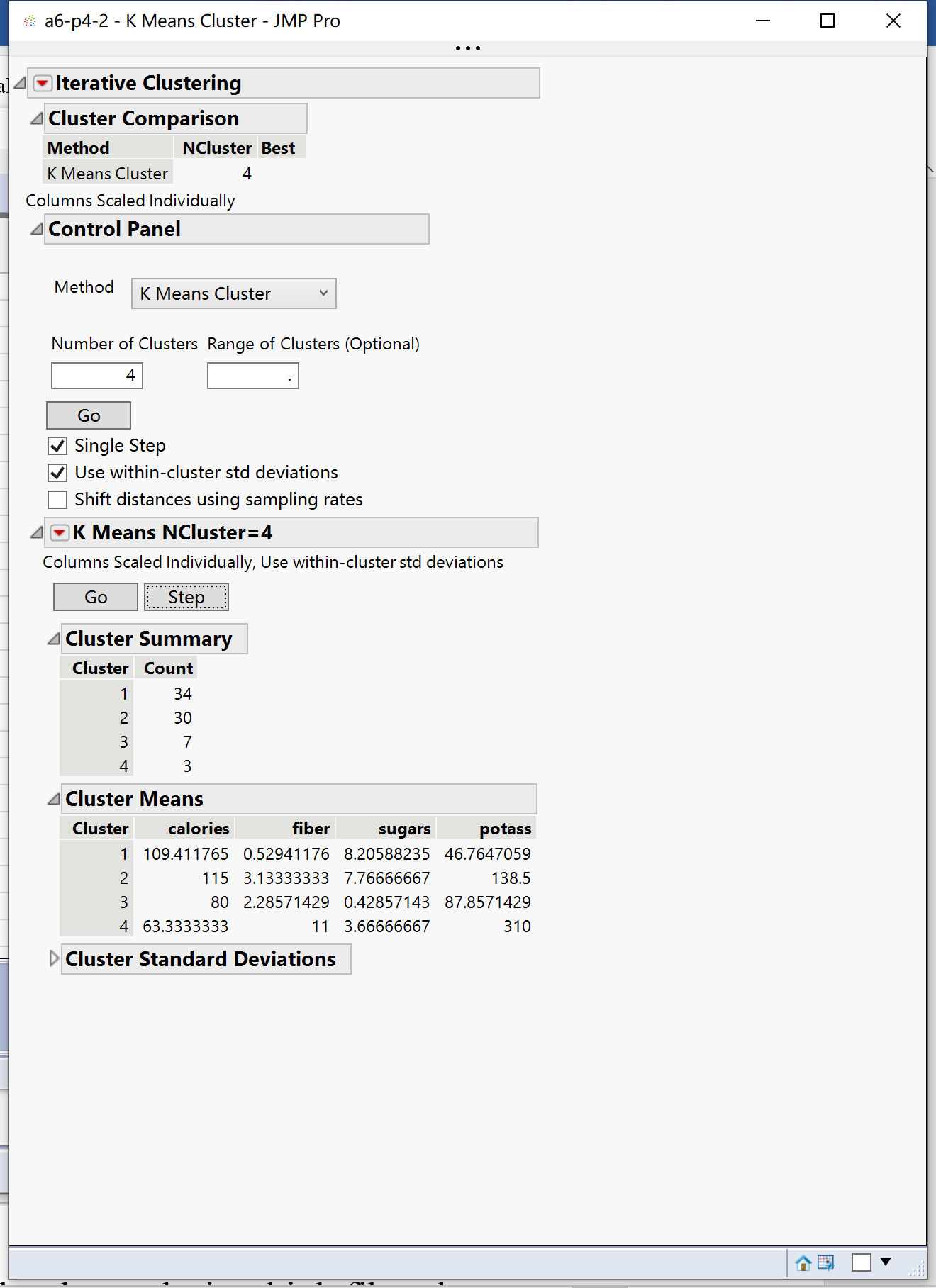
I use JMP, and the screenshot are shown below.

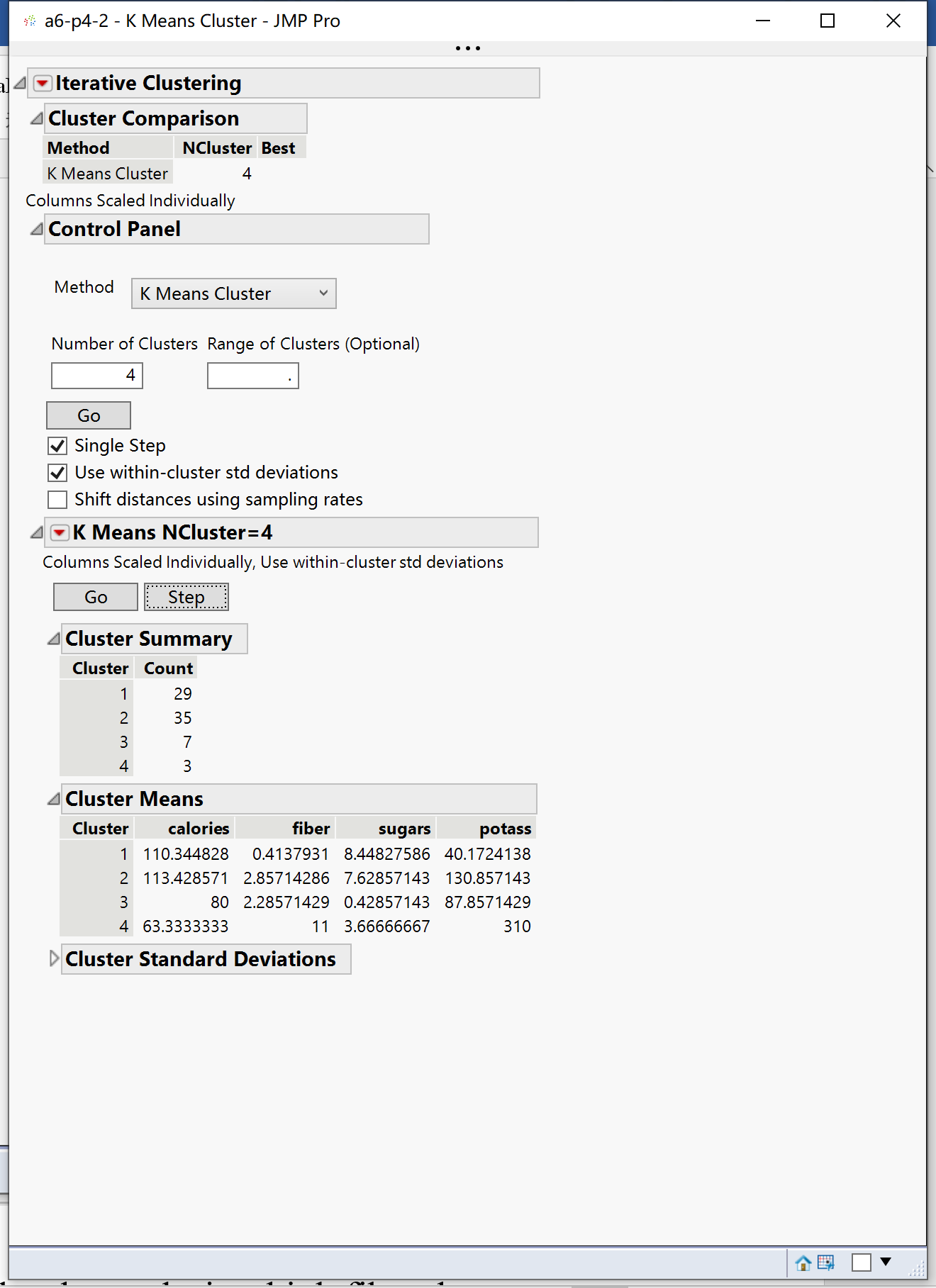
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | cluster | | | |
| 1 | 2 | 3 | 4 |
| C a l o r i e s m e a n | 110.344828 | 113.428571 | 80 | 63.33333 |
| stddev | 4.90092083 | 18.6613209 | 20.7019668 | 9.42809042 |
| max | 110.344828 | 128 | 80 | 63.33333 |
| min | 107.291667 | 113.428571 | 50 | 50 |
| F i b e r mean | 0.4137931 | 2.85714286 | 2.28571429 | 11 |
| stddev | 0.49251231 | 1.26813377 | 1.27775313 | 2.1602469 |
| max | 1.29090909 | 6 | 2.33333333 | 14 |
| min | 0 | 2.85714286 | 1 | 11 |
| S u g a r s mean | 8.44827586 | 7.62857143 | 0.42857143 | 3.666667 |
| stddev | 4.50735955 | 3.22566241 | 0.72843136 | 2.62466929 |
| max | 12 | 14 | 0.42857143 | 3.666667 |
| min | 7.25 | 7.62857143 | 0 | 0 |
| Potassium mean | 40.1724138 | 130.857143 | 87.8571429 | 310 |
| stddev | 12.3513756 | 47.7416517 | 39.357078 | 21.602469 |
| max | 70.5454545 | 260 | 87.8571429 | 330 |
| min | 25 | 130.857143 | 50 | 310 |









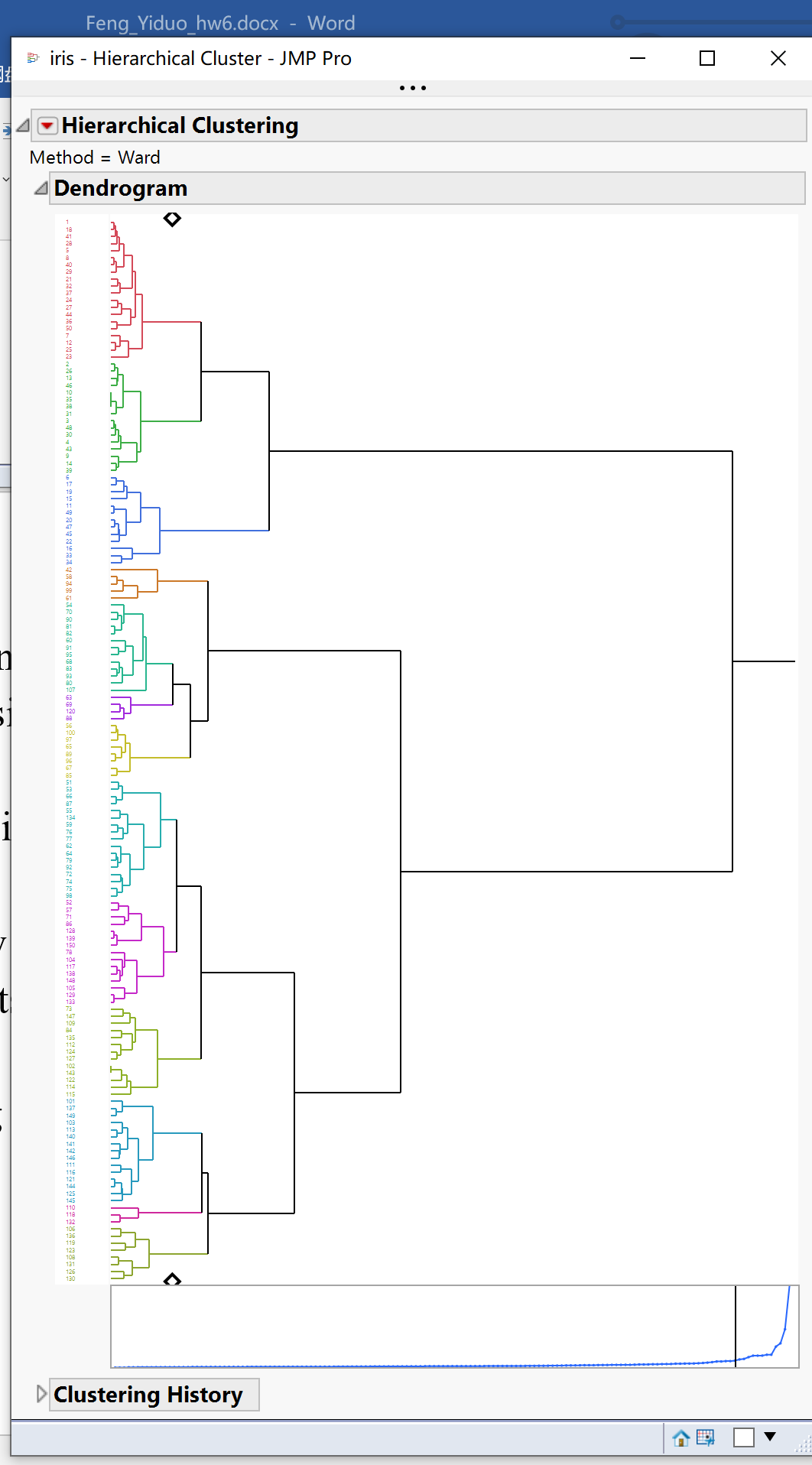


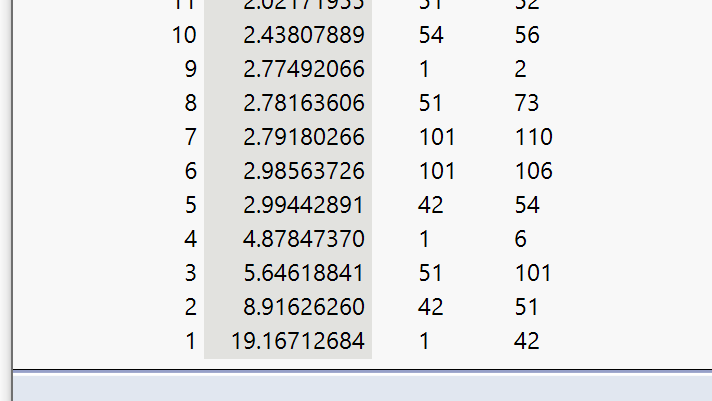
(2)-b. In general, a cereal is considered healthy if it has low calories, high fiber, low sugars, and high potassium. Using this standard, which cluster has the healthiest cereals?

According the table above, using this standard, 4 cluster has the healthiest cereals.

**Problem 5 (20 points).** Follow the instructions in *JMP-clustering-assignment.pdf* file. Include the required screenshots and your answers to some questions in your submission.

1. Hierarchical Clustering



According to the elbow method, 5 is the optimal number. From 1 to 4, the distance changed a lot and not stable until it gets 5. The numbers between the distances of 5 and 6 is only 0.01, and from 5 to 10, the distances don’t change a lot.

1. K Means

