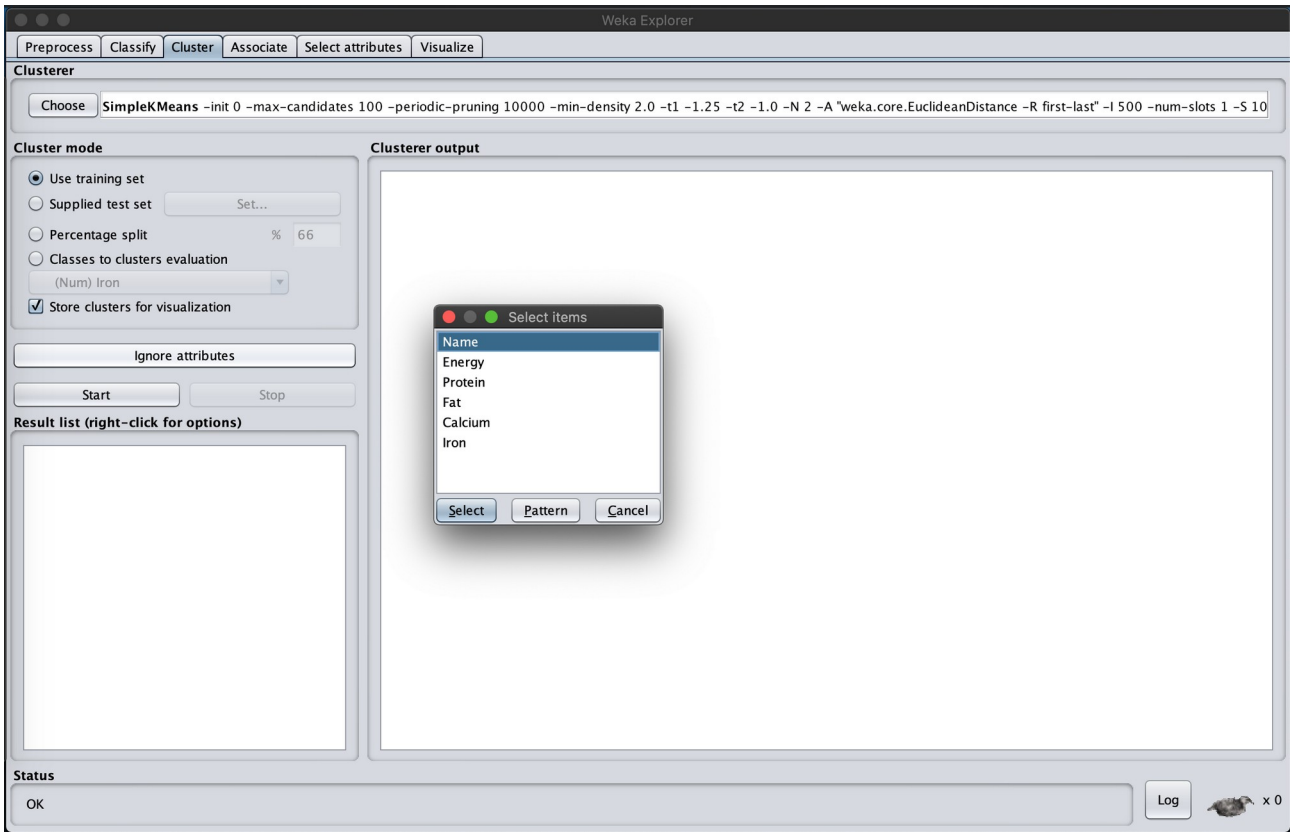
Data Mining Lab1-Clustering

Zhixuan_Duan(zhidu838)
2020-02-10

1. Simple Kmeans

Apply "Simple KMeans" to your data. In the Weka, Euclidian distance is implemented in SimpleKmeans. You can set the number of clusters and seed of a random algorithm for generating initial cluster centers. Experiment with the algorithm as follows:

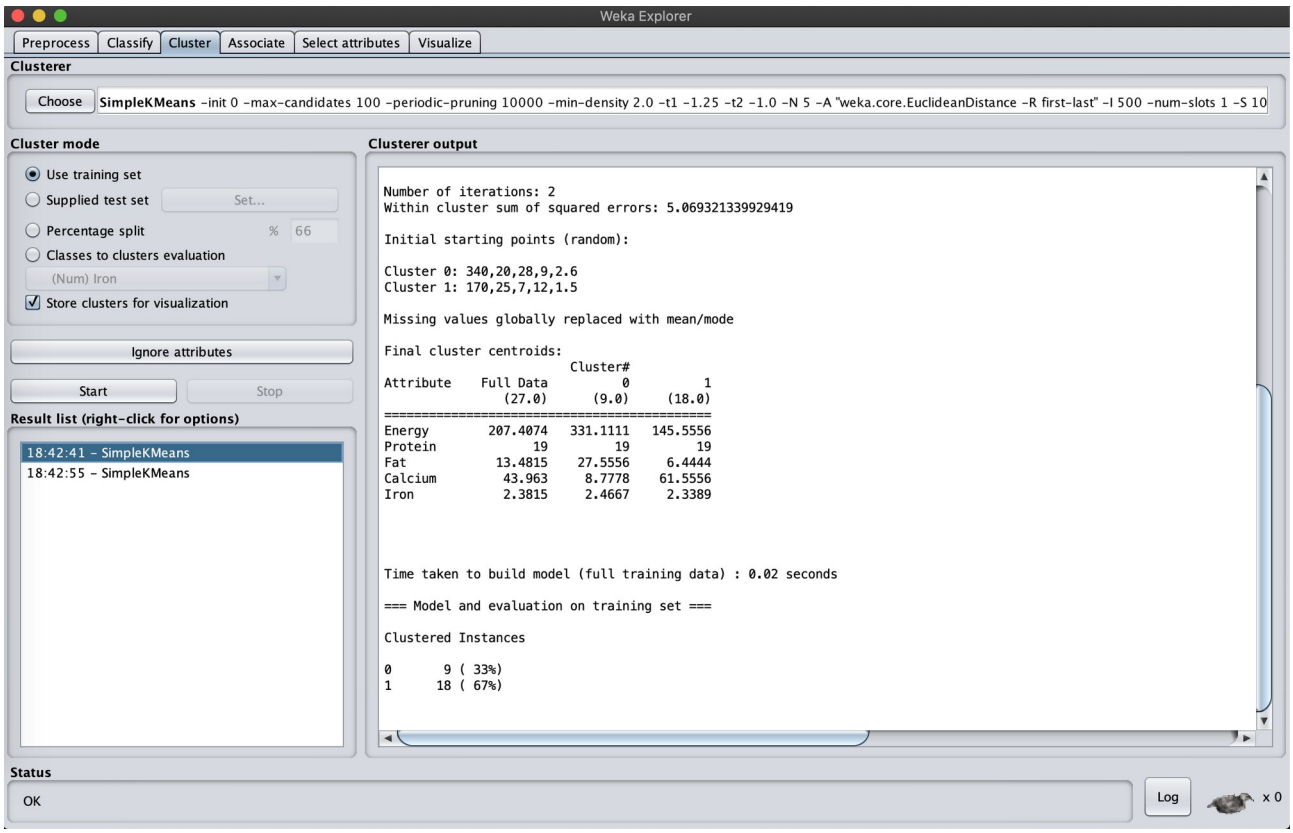
1. Choose a set of attributes for clustering and give a motivation. (Hint: always ignore attribute "name". Why does the name attribute need to be ignored?)

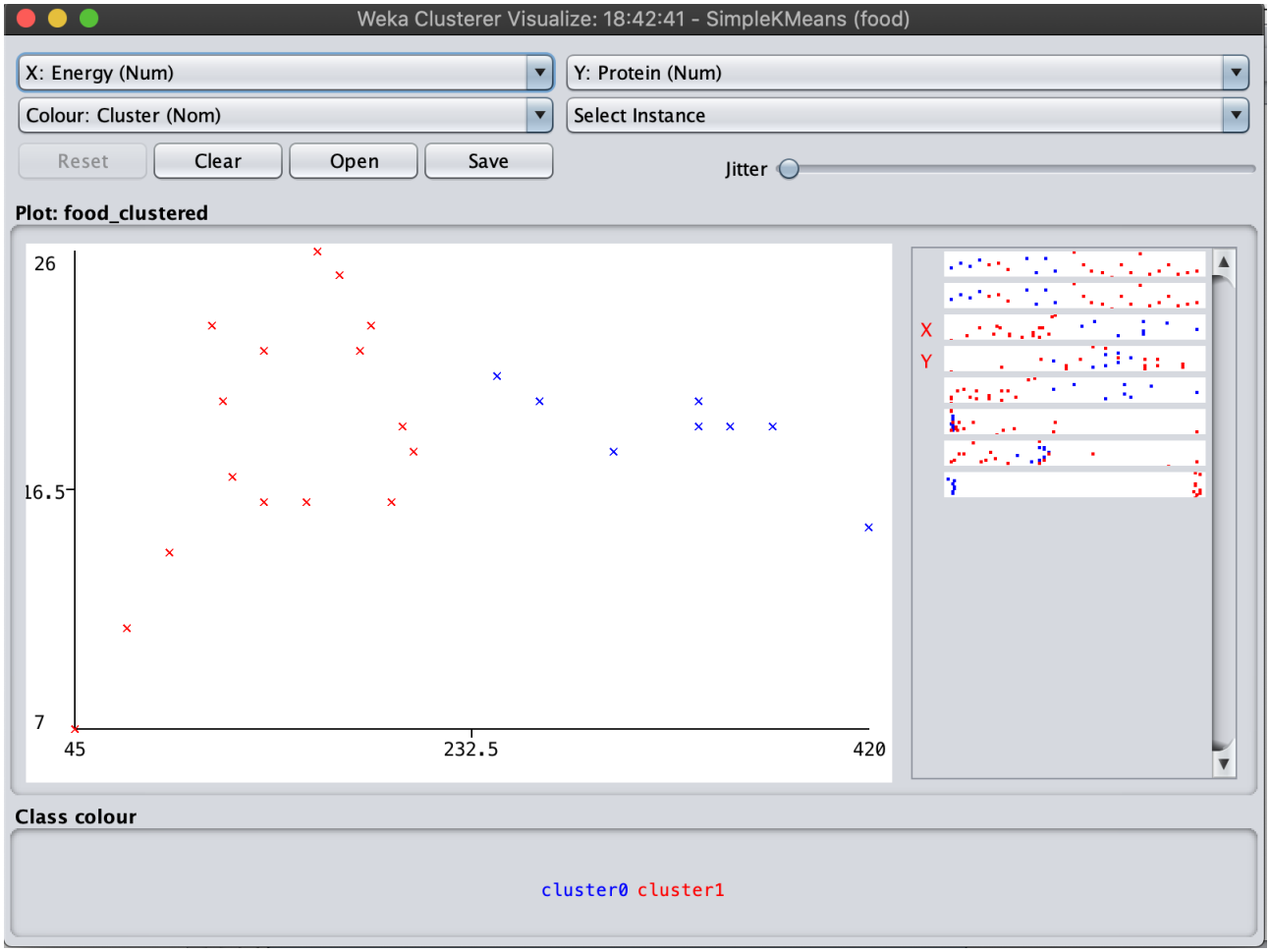


Analysis:
Choose attributes except name for clustering to analysis if there is hidden relationship between observations. Like if the input of energy related to input of fat. By clustering, we may find out some knowledge by this unsupervised method.

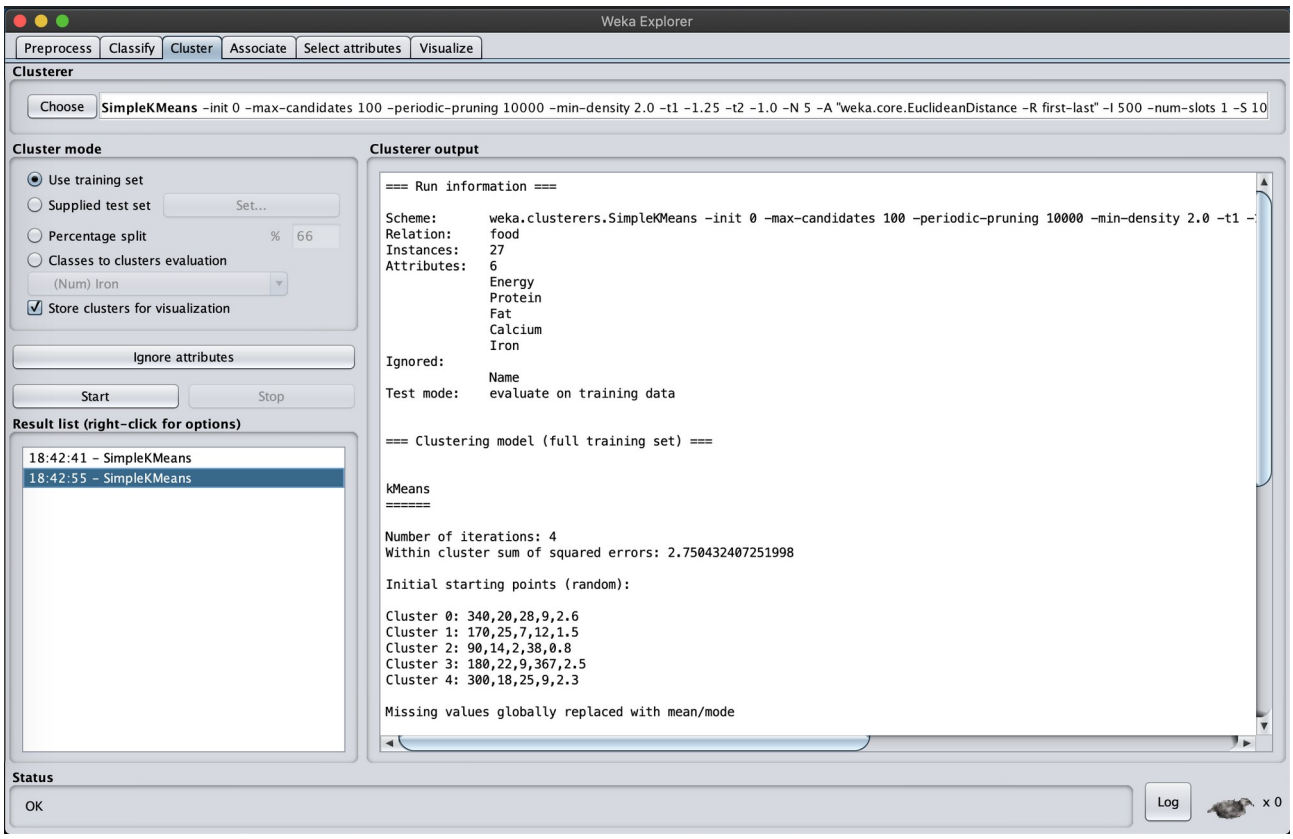
Because the clustering method are normally used to deal with numerical data, which are relatively easy to calculate the distance between each node, however, the attribute “Name” is a character variable, so the distance may be disturbed.

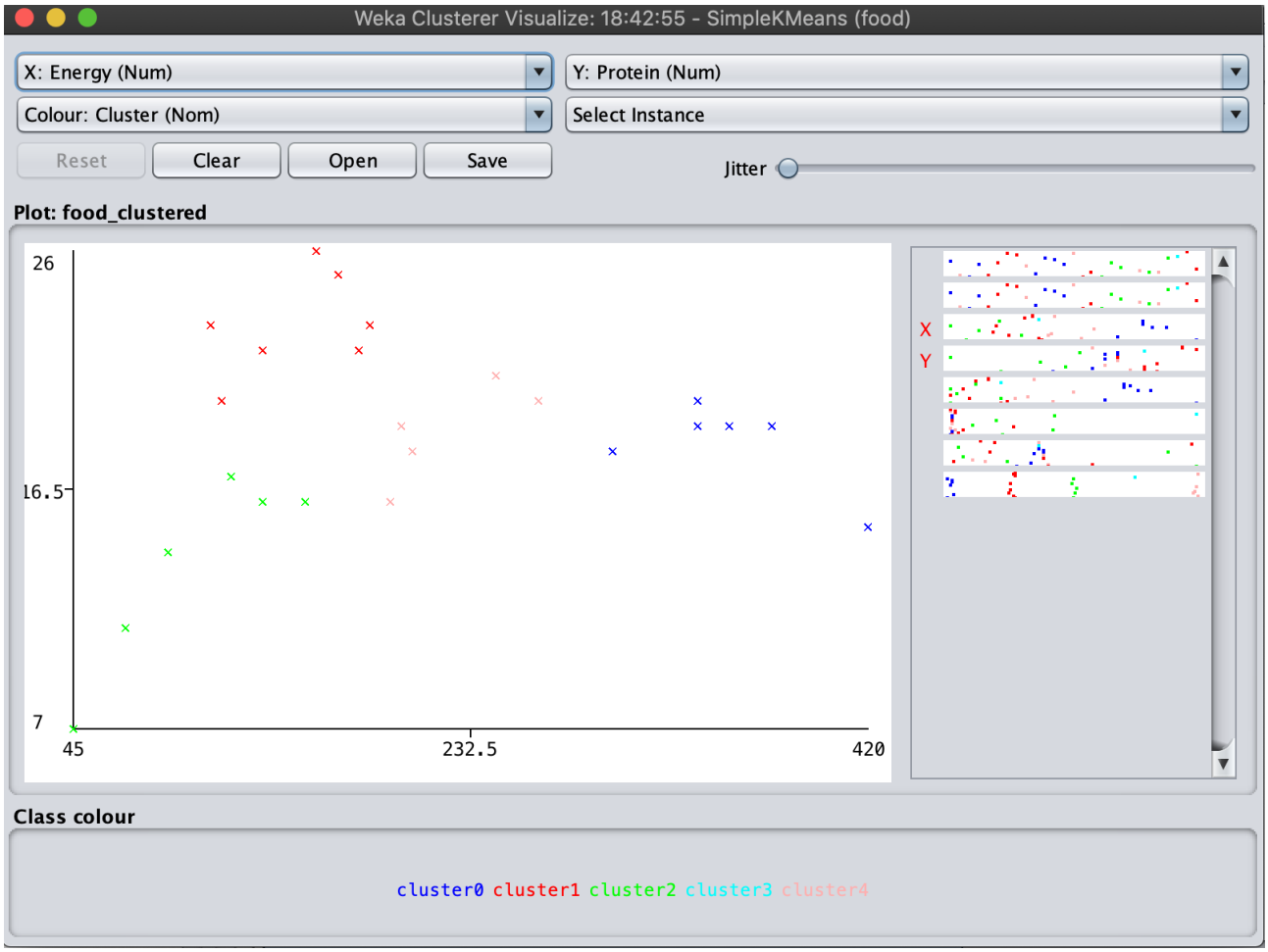
2. Experiment with at least two different numbers of clusters, e.g. 2 and 5, but with the same seed value 10.





The above results uses 2 clusters.



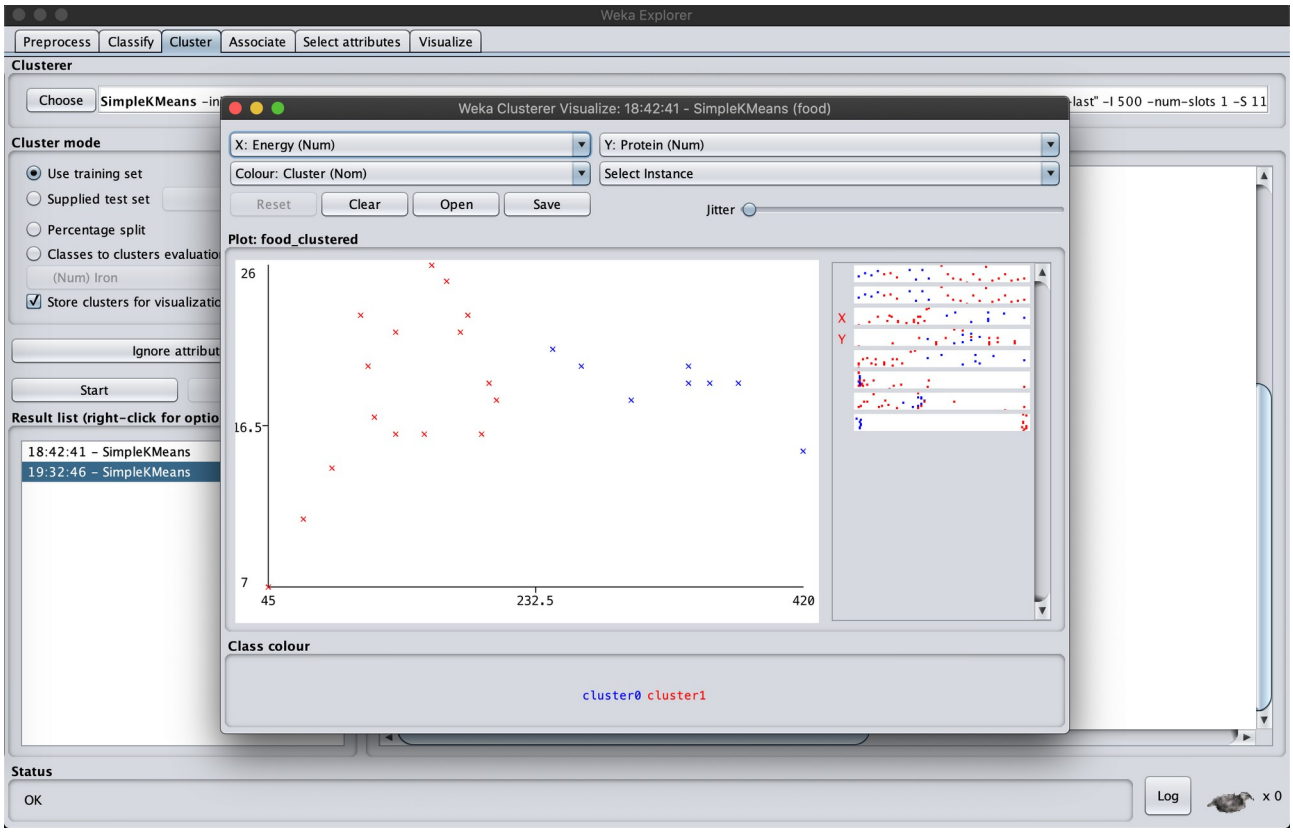


The above results uses 5 clusters.

Analysis:
Compare these two results, since the simple k-means method would not produce any overlap, so both results seems reasonable under different situations. So it needs advanced analysis to choose between different clusters.

And if we check the output results, we can see that the k-means method cannot avoid the impact of extreme values, since in the 2 clustering results, some of blue points are forced to be separated and establish a new clustering in 5 clustering condition since there is a extreme value of (420, 15), which influence the distance calculation.

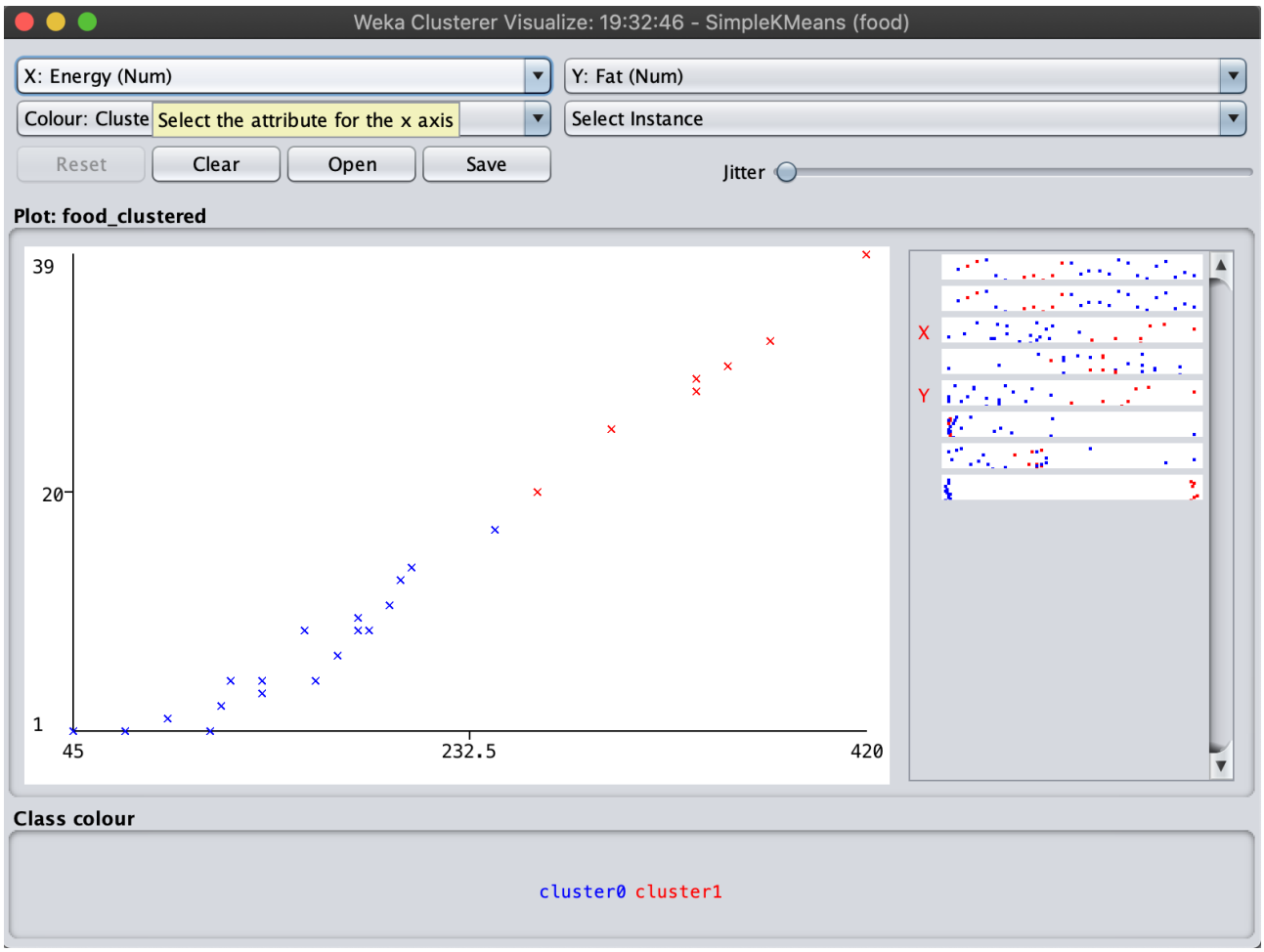
3. Then try with a different seed value, i.e. different initial cluster centers. Compare the results with the previous results. Explain what the seed value controls.



Analysis:
This time we choose seed as 11, so from the results, we can see that different seed value change the initial node of cluster centroid.

And if we compare with previous clustering, we can see when number of clusters is relative small, changing seed does not change the results a lot, but when number of clustering increases, the change of seed number would become more unstable, since the distance calculation would be totally different.

4. Do you think the clusters are "good" clusters? (Are all of its members "similar" to each other? Are members from different clusters dissimilar?)

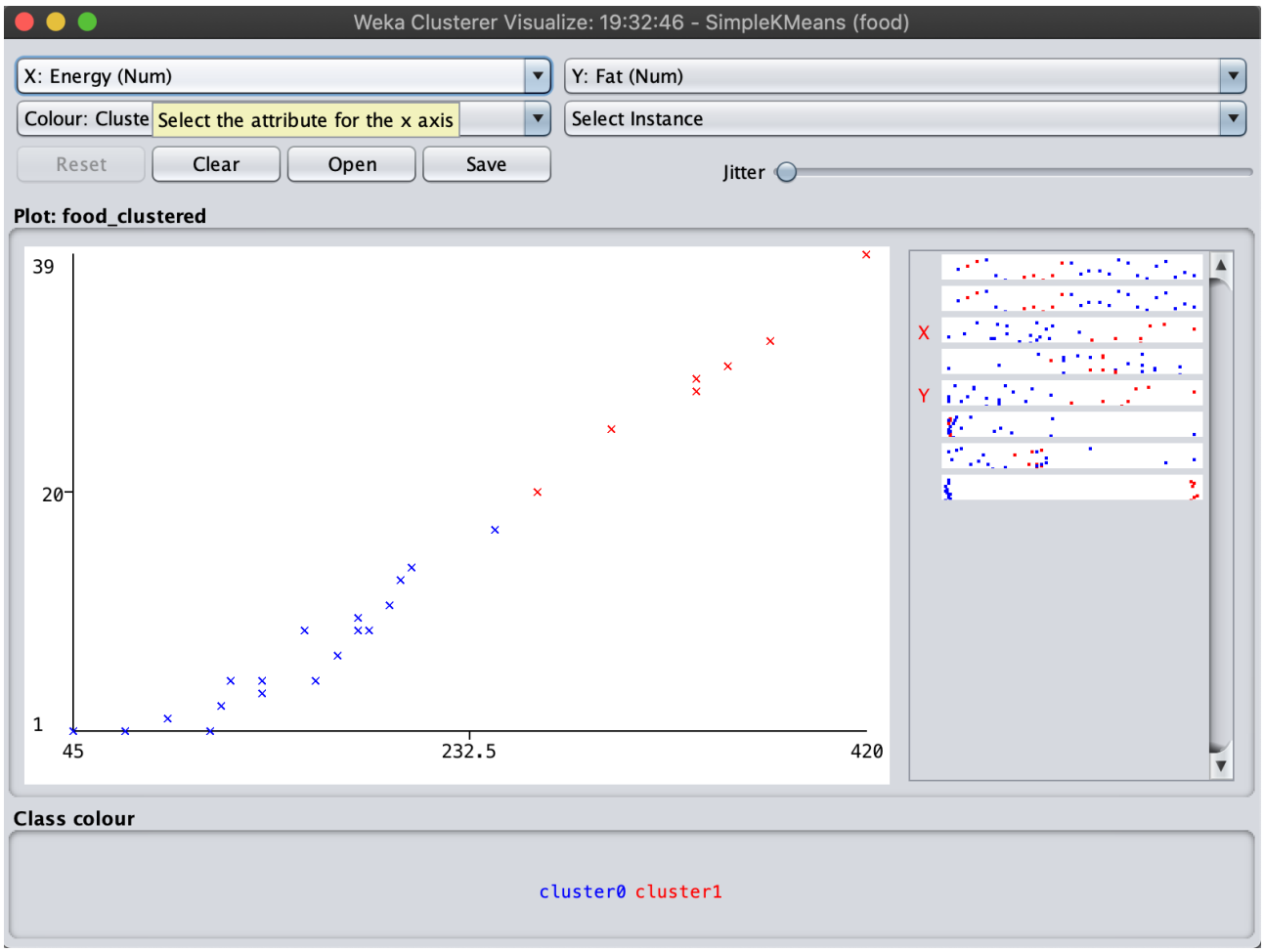


Analysis:
There is no background material provided, so the following analysis is solely based on the method itself and the output result.

From the plots above, we can see that the Within cluster sum of squared errors got the least to 5.08, which is quite small. The errors between clusters are much bigger than it, so I think it is a good clustering.

Furthermore, from the visualizations, members from different clusters are quite dissimilar, which also illustrates the point made above.

5. What does each cluster represent? Choose one of the results. Make up labels (words or phrases in English) which characterize each cluster.

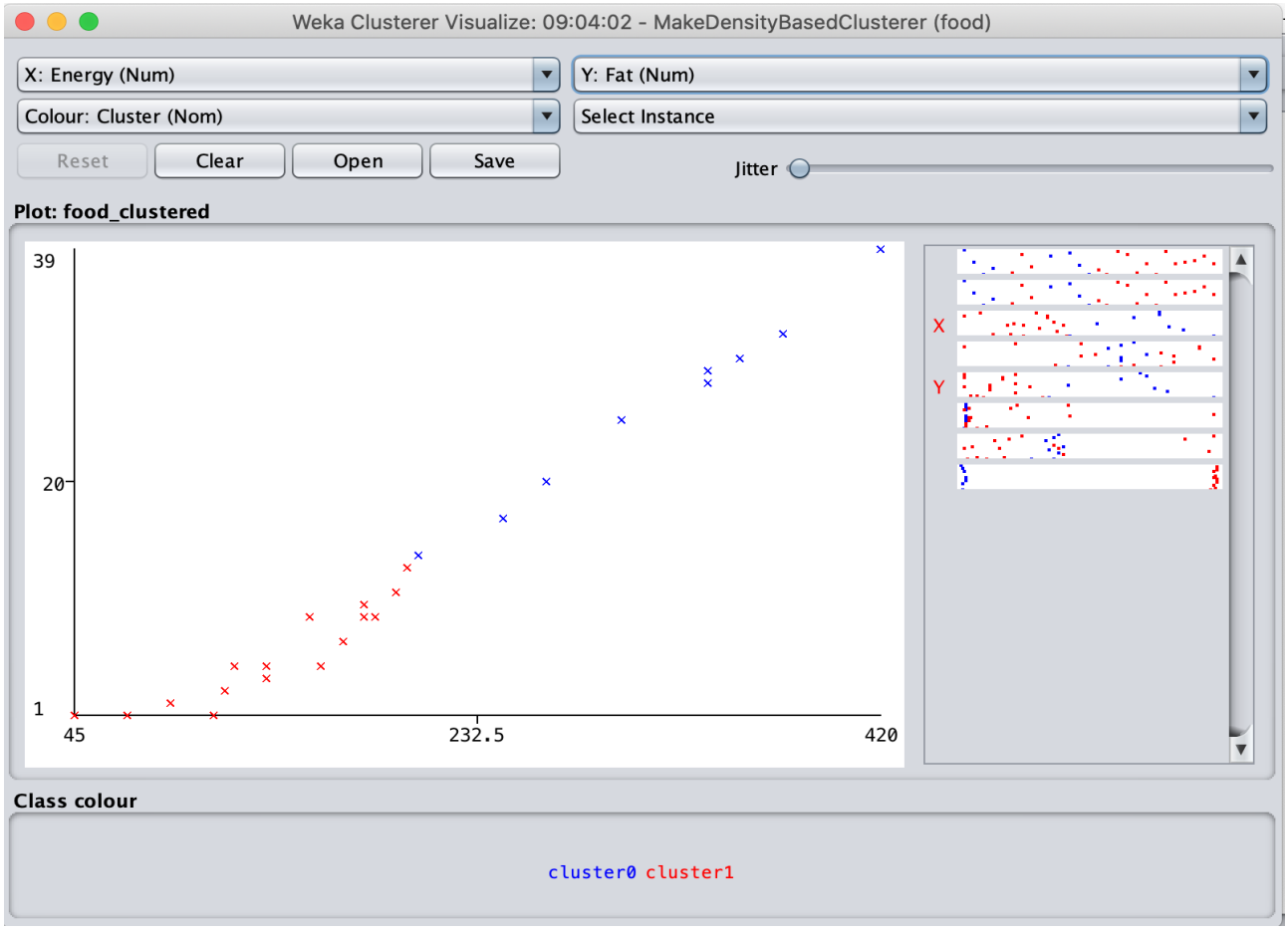


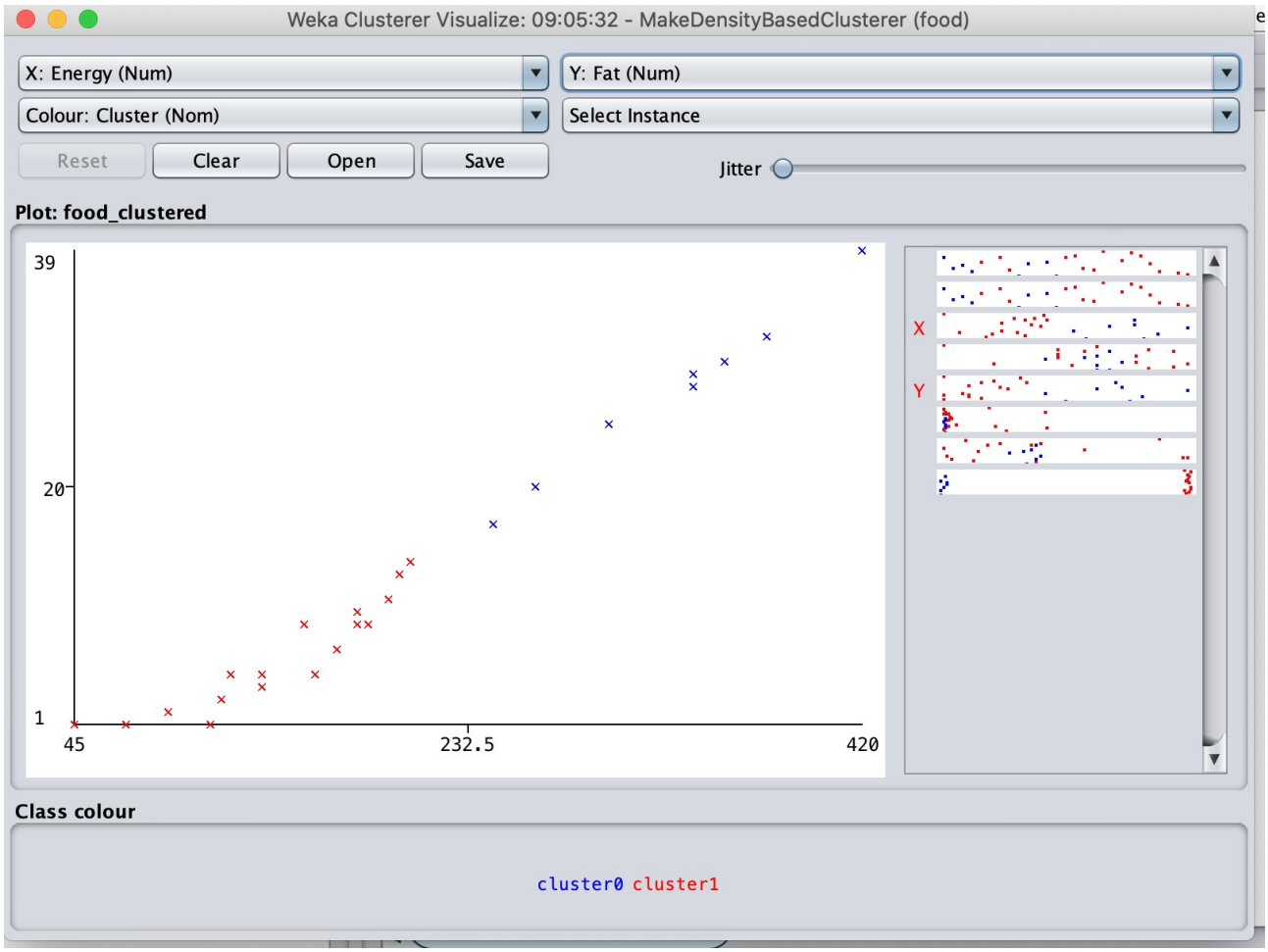
Analysis:
From this clustering, we choose the Energy and the Fat as input, set k to 2.

We can see from the plot that lower energy relates to lower fat in the blue cluster, and higher values occurred in both cases for red cluster. It makes sense because the input of energy including fat, so natural there is a positive relationship between these two attributes.

2. MakeDensityBasedClusters
Now with MakeDensityBasedClusters, SimpleKMeans is turned into a denstiy-based clusterer. You can set the minimum standard deviation for normal density calculation. Experiment with the algorithm as the follows:

1. Use the SimpleKMeans clusterer which gave the result you haven chosen in 5).
2. Experiment with at least two different standard deviations. Compare the results. (Hint: Increasing the standard deviation to higher values will make the differences in different runs more obvious and thus it will be easier to conclude what the parameter does)


Weka Clusterer Visualize: 09:04:02 - MakeDensityBasedClusterer (food)
X: Energy (Num)
Y: Fat (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: food_clustered
39
20
1
45
232.5
420
X
Y
Class colour
cluster0 cluster1


Weka Clusterer Visualize: 09:05:32 - MakeDensityBasedClusterer (food)
X: Energy (Num)
Y: Fat (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: food_clustered
39
20
1
45
232.5
420
X
Y
Class colour
cluster0 cluster1

Analysis:
We choose the same attributes Energy and Fat as in the 1 5), and choose Density-based clustering method.

The first clustering is done by using default sd as (1e-6), and the second one is got by setting sd to 10.

Compare the first result with which in 1 5), we can see density-based method produce different output, that some of observations are set to red cluster, because in this method, it calculates the distance based on density rather than means.

And compare the results under different sd value, we can see that the standard deviation controls accept/reject new nodes into original clusters, if the sd is too large, each cluster would has higher tolerance to new nodes, and in order to satisfy the sd condition, it will sacrifice similarity within each cluster.