732A75 Data Mining Lab-1

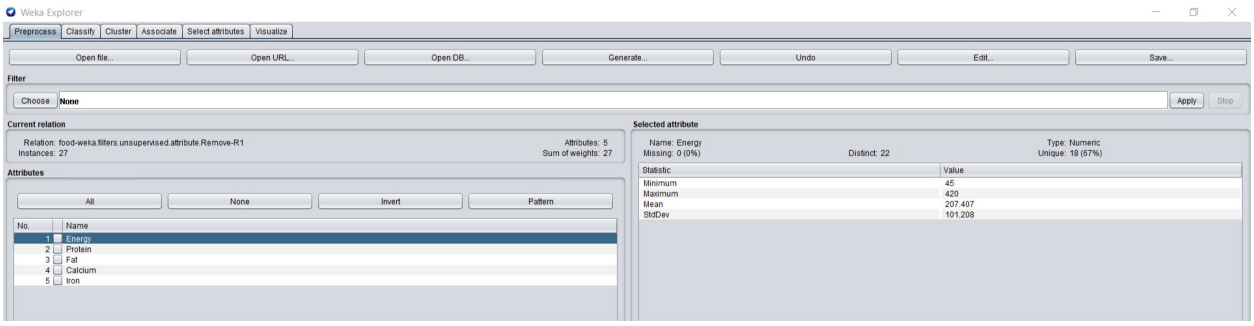
Sridhar Adhikarla(sriad858) and Lakshidaa Saigiridharan(laksa656)

5 March 2019

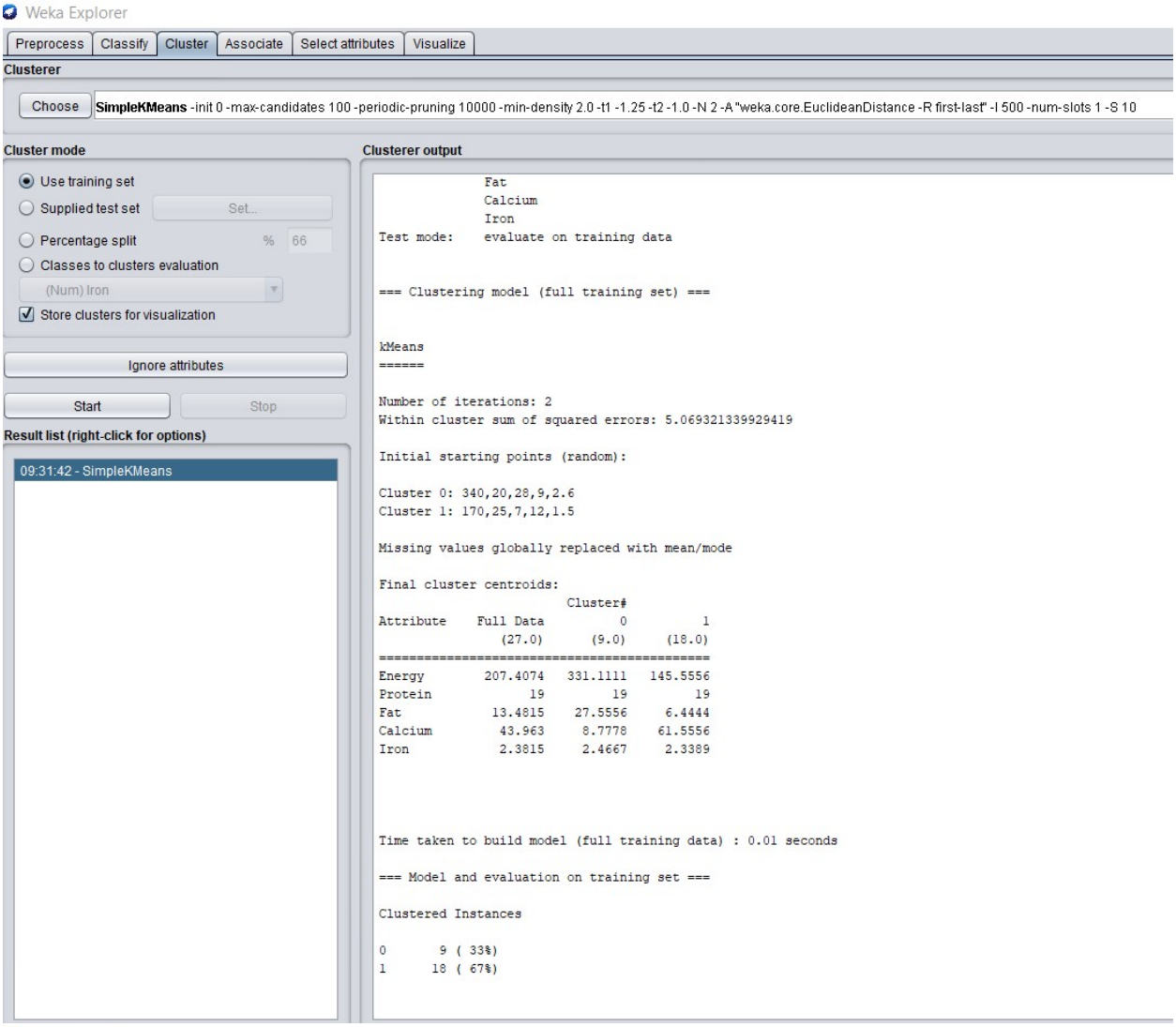
SimpleKmeans:

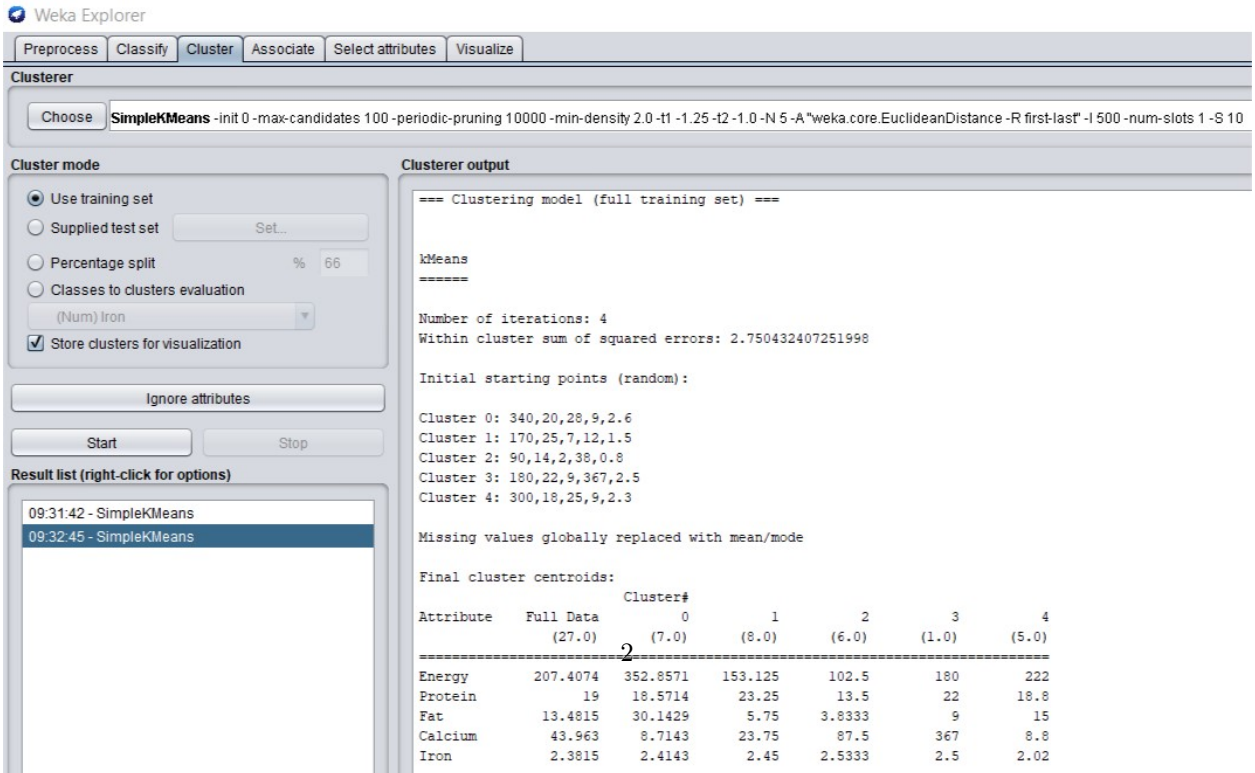
1. Choose a set of attributes for clustering and give a motivation.

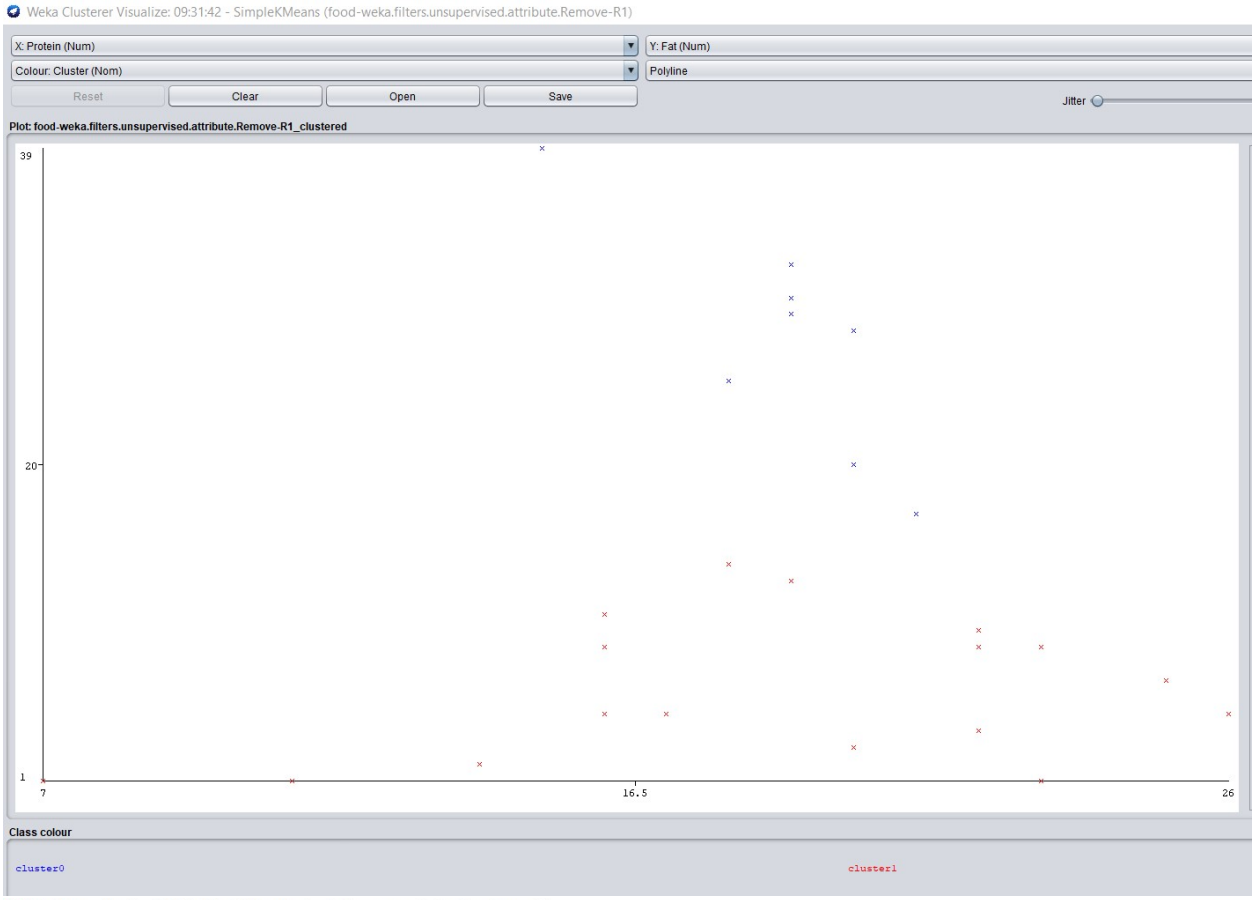
Usually in clustering operations we only use numeric attributes, any attributes of categorical type or having too many levels are excluded. It is due to the fact that clustering tends to be based on some distance measurment, and distance of attributes such as name, rownumbers are meaningless and may only come in the way of optimal solution.

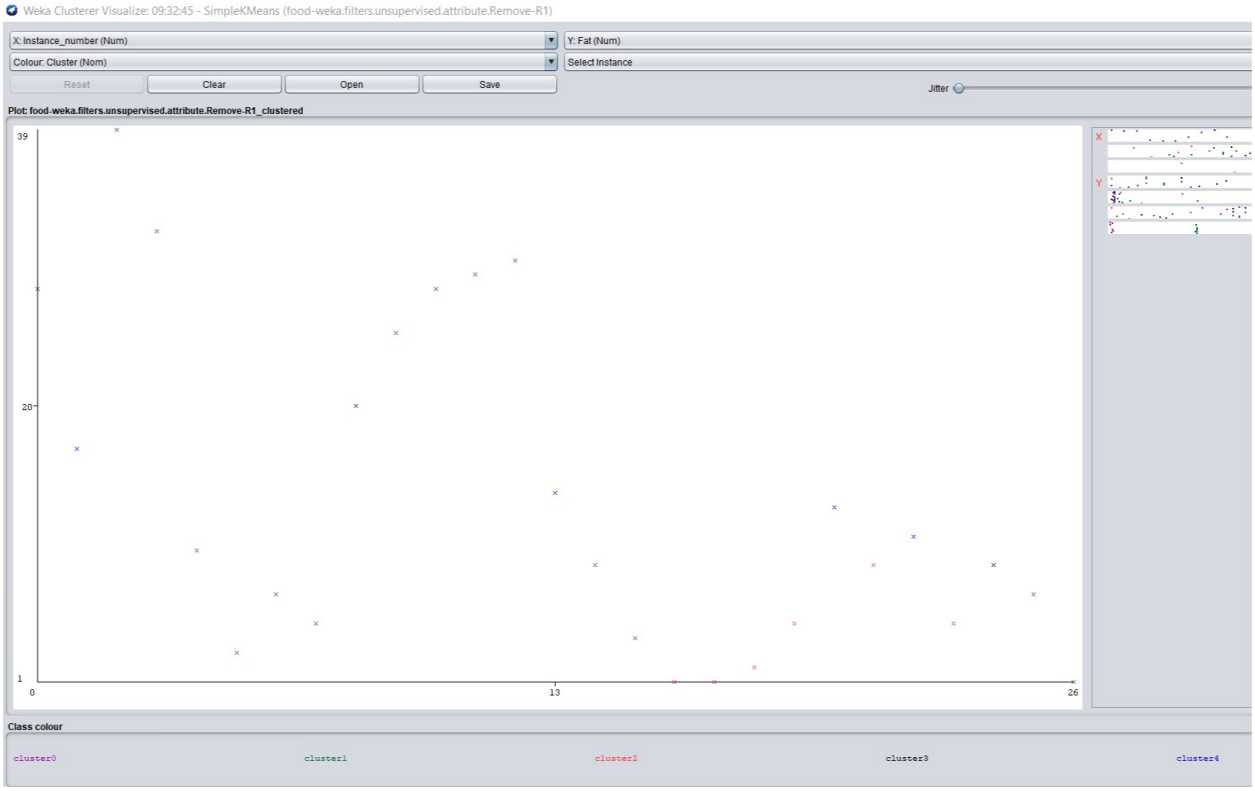


2. Experiment with at least two different numbers of clusters.



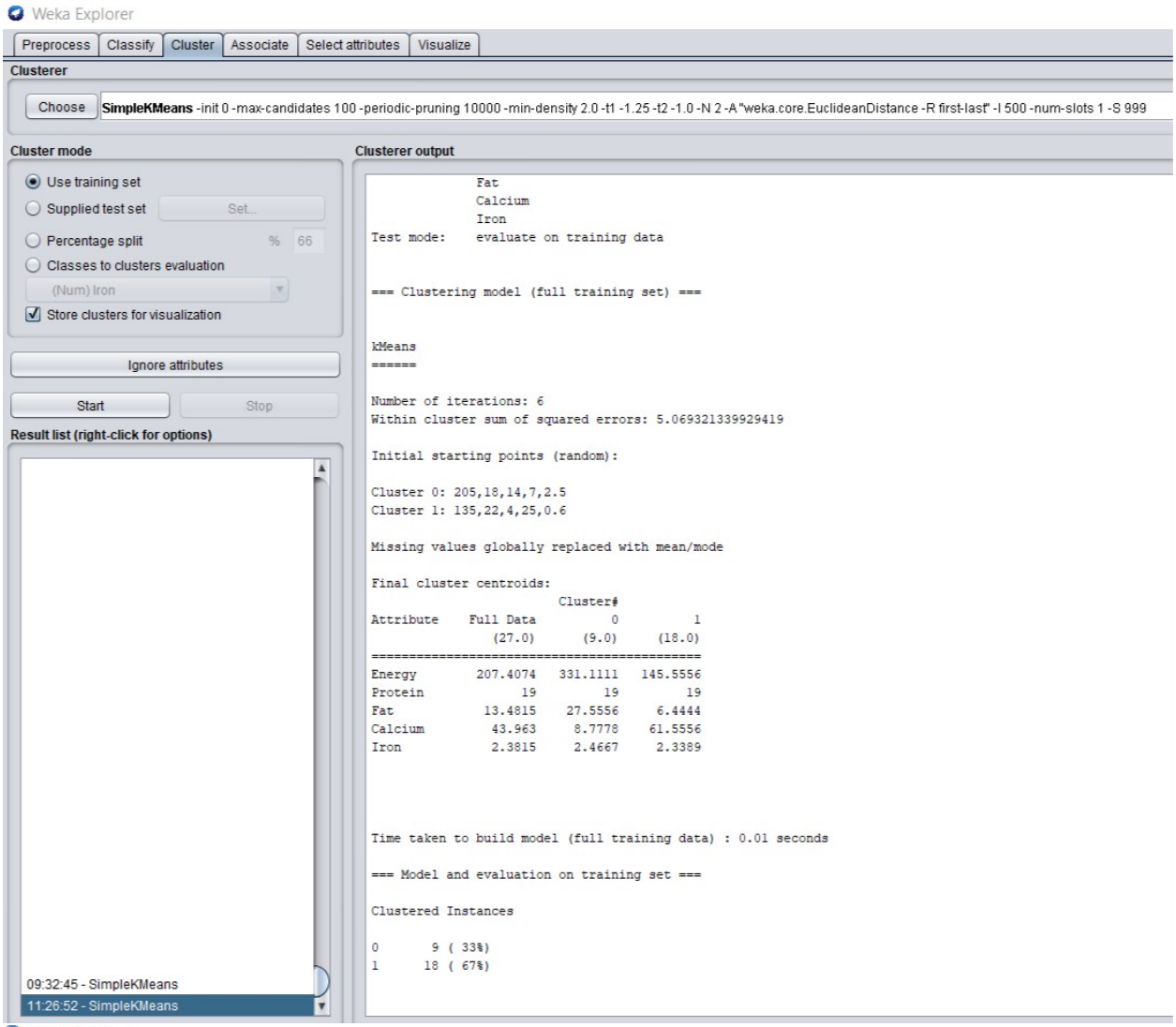


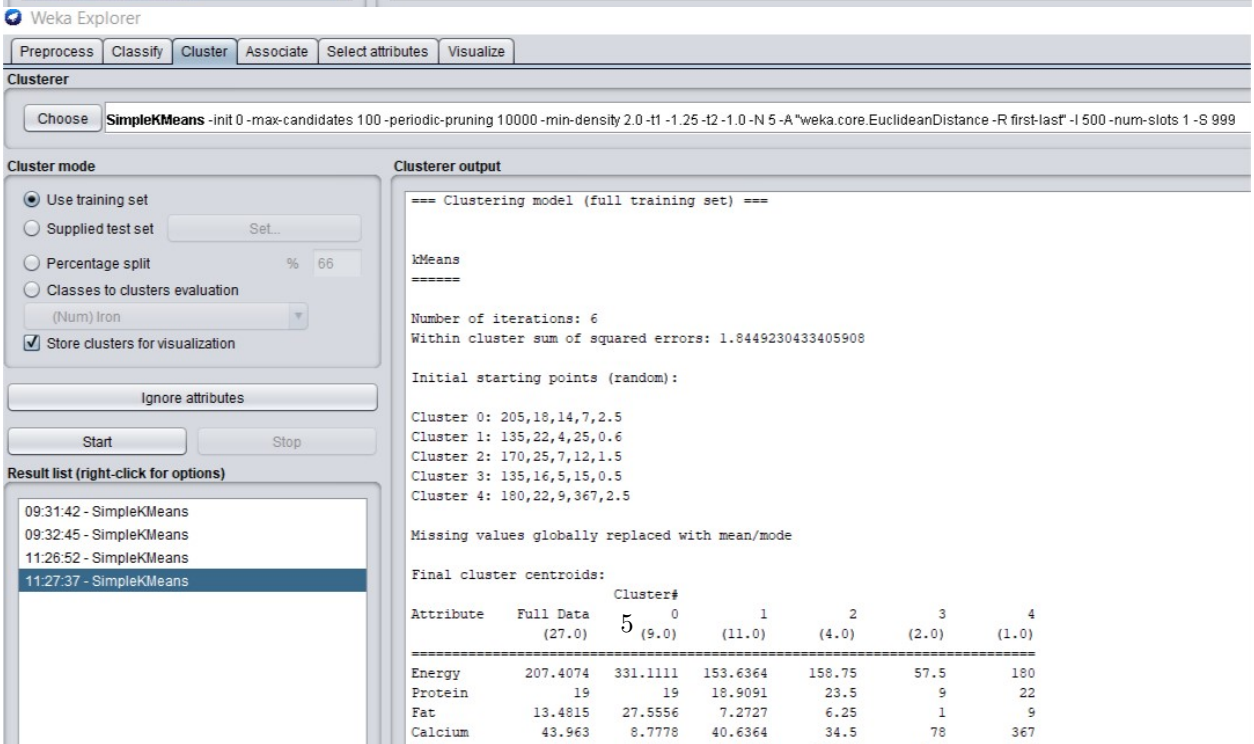

Weka Clusterer Visualize: 09:31:42 - SimpleKMeans (food-weka.filters.unsupervised.attribute.Remove-R1)
X: Protein (Num)
Y: Fat (Num)
Colour: Cluster (Nom)
Polyline
Reset
Clear
Open
Save
Jitter
Plot: food-weka.filters.unsupervised.attribute.Remove-R1_clustered
39
20
1
7
16.5
26
Class colour
cluster0
cluster1

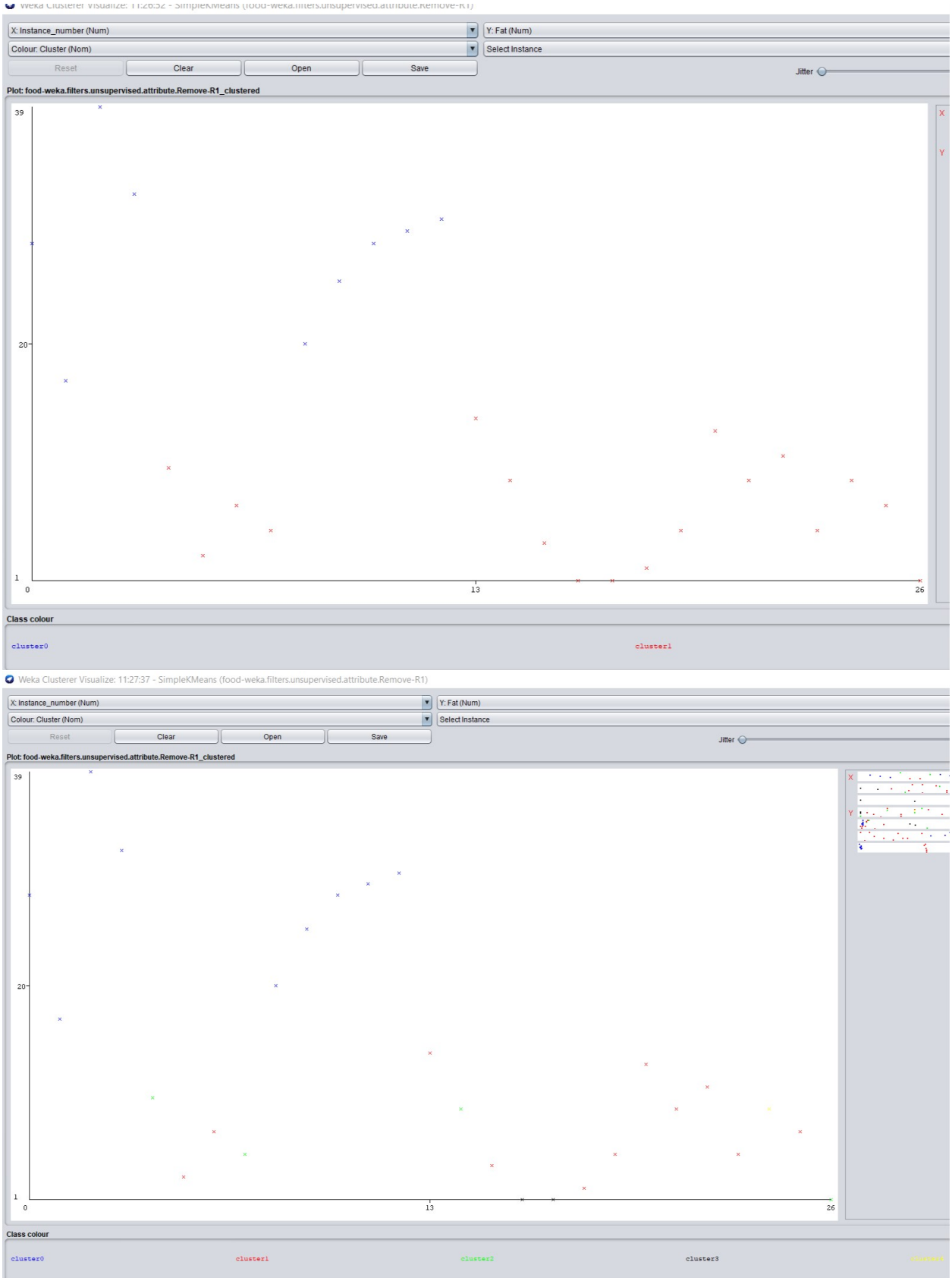

Weka Clusterer Visualize: 09:32:45 - SimpleKMeans (food-weka.filters.unsupervised.attribute.Remove-R1)
X: Instance_number (Num)
Y: Fat (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: food-weka.filters.unsupervised.attribute.Remove-R1_clustered
39
20
1
0
13
26
Class colour
cluster0
cluster1
cluster2
cluster3
cluster4

K-means is a partion based clustering algorithm in which as we increase the number of cluster, the number of subdivisions increase. It is due to the fact that none of the clusters can be empty or can overlap in a partion based clustering algorithm, and it covers all the data points. This can lead to the possibility that some clusters have just one or two points, which is meaningless. Thus care must be taken to ensure that we specify k carefully and try and inspect visually to see if the clusters make sense.

3. Try with a different seed value. Compare the results with the previous results. Explain what the seed value controls.



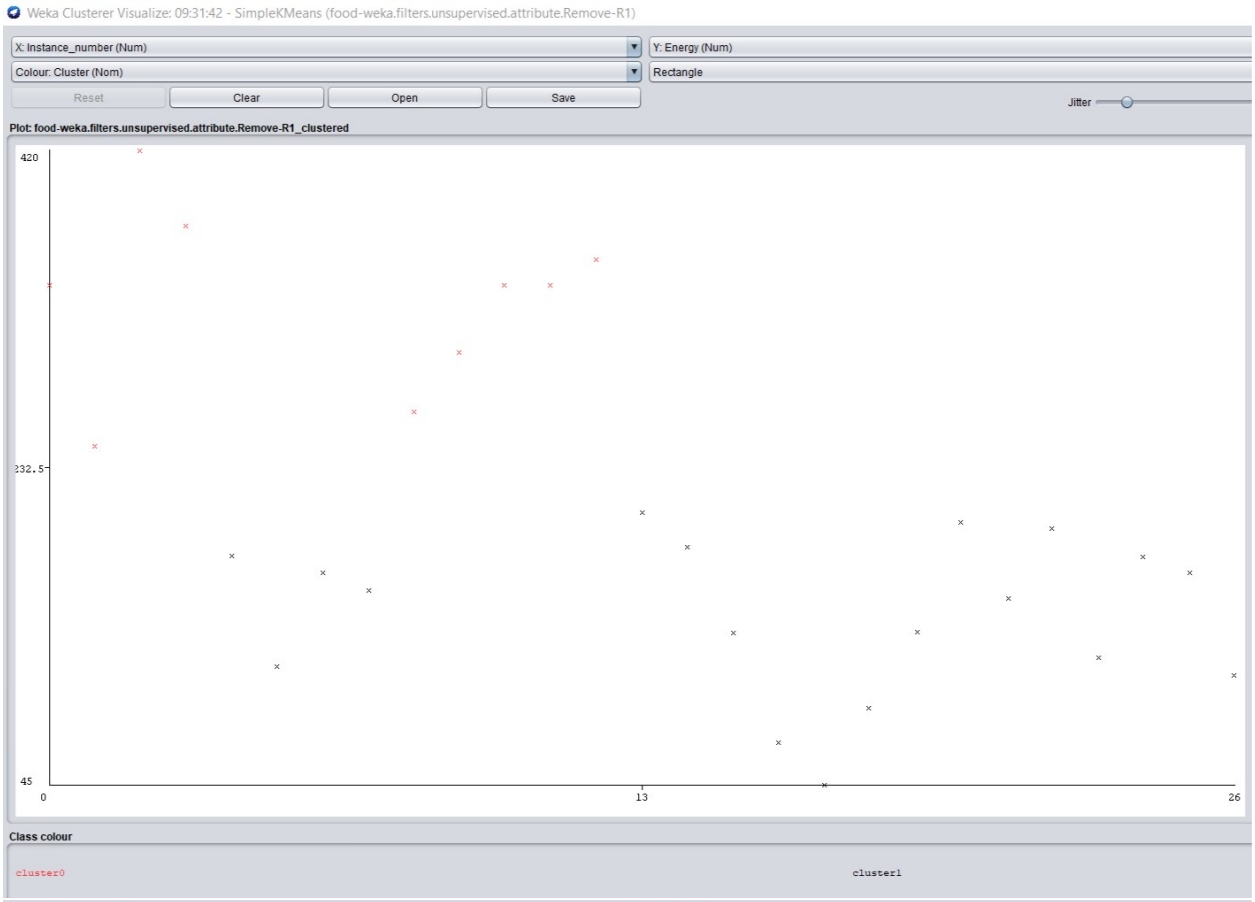


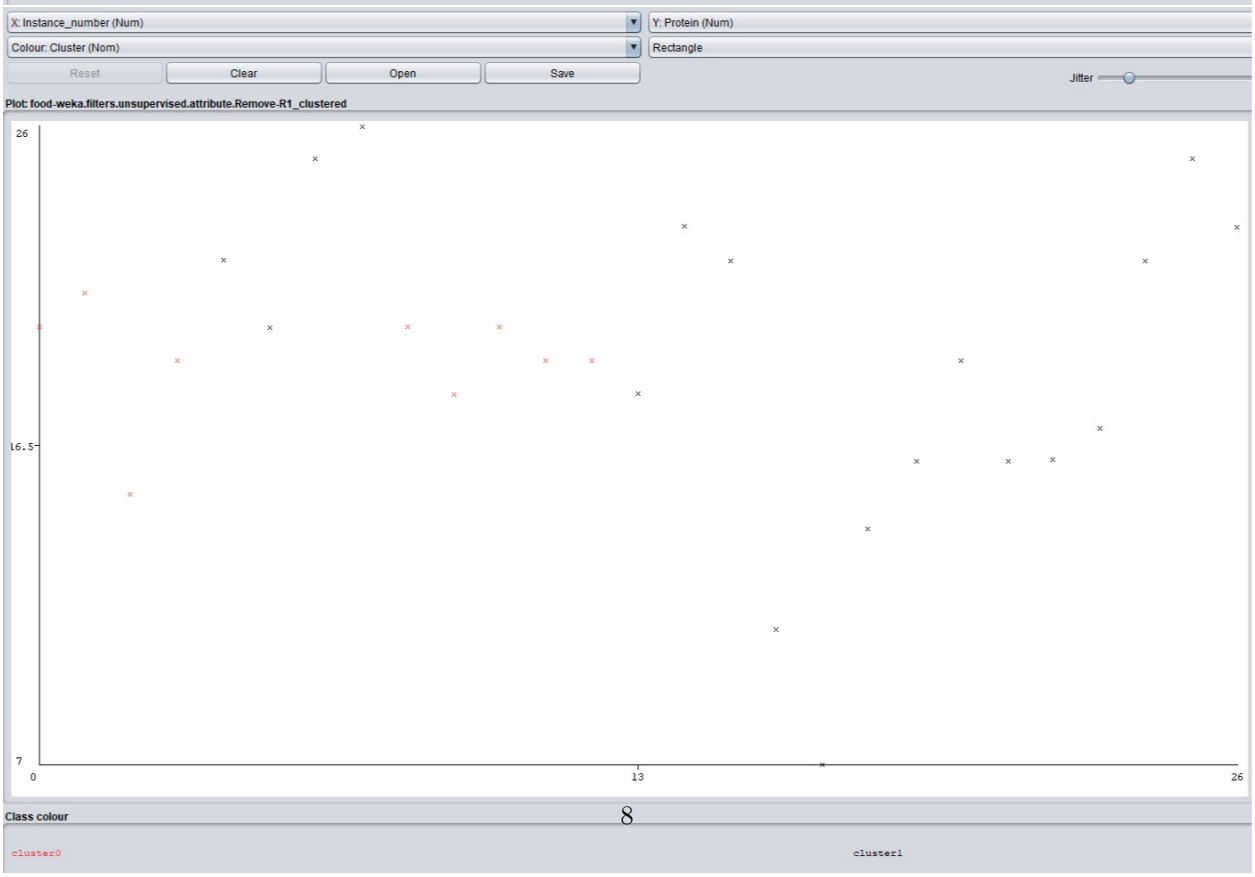


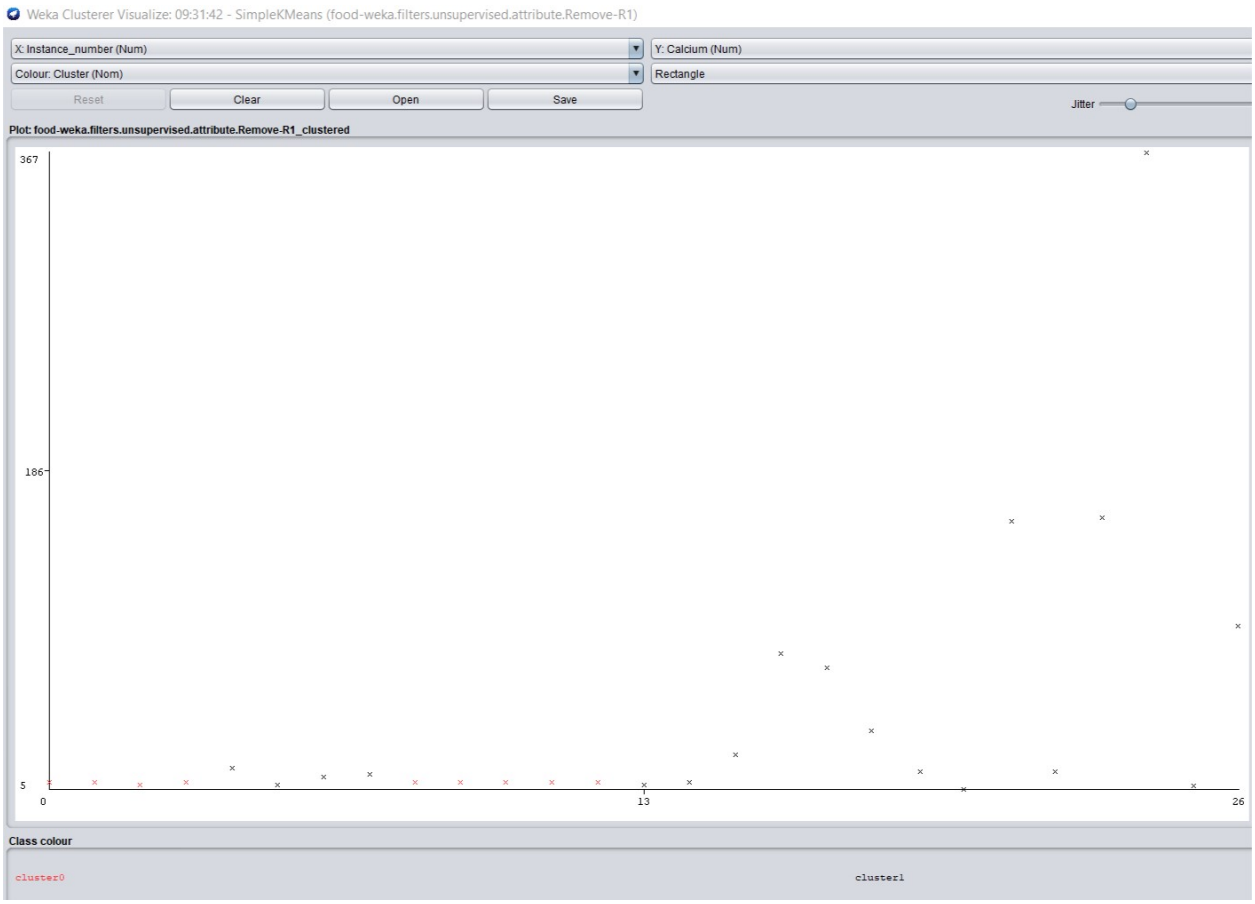
The seed value controls the starting points of the cluster centroids. If we change the value of the seed, the

algorithm seems much more unstable for larger values of k, producing quite different clusters. This could be due to the fact that we have very few data points. However, for smallr k values, the clusters produced are quite similar for different seeds. With $k = 2$ for instance, the results are quite similar, evaluating by the centroids of the clusters, however for $k = 5$, much less so. We chose these values in order to compare how meaningful the clusters are when there are many vs when there are few. The number of data points in the clusters may vary too with the seed.

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



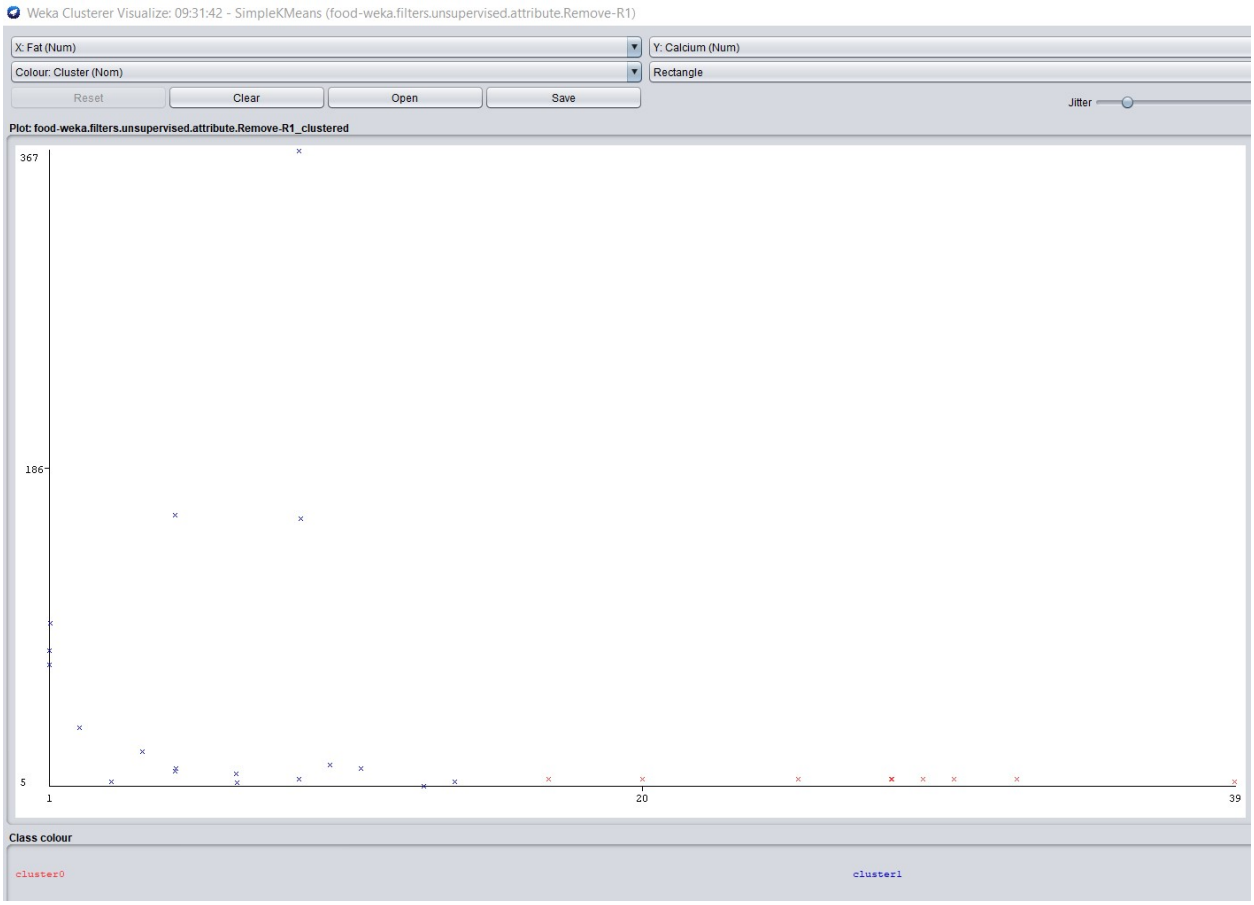




We evaluate the goodness of a cluster based on the attributes we are looking at. For more than two attributes it is hard to visualize them, this is the reason we visualize the cluster against each of the attributes to evaluate them.

The clusters seperate some attributes(Energy) very well, but not so good with other attributes(Calcium). In a well done cluster the elements wethin a cluster are closer to each other and the cluster are seperated(further the better). We can see that in the plot of cluster vs Energy. The red cluster elements are a little closely packed and well seperated from the blue cluster.

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

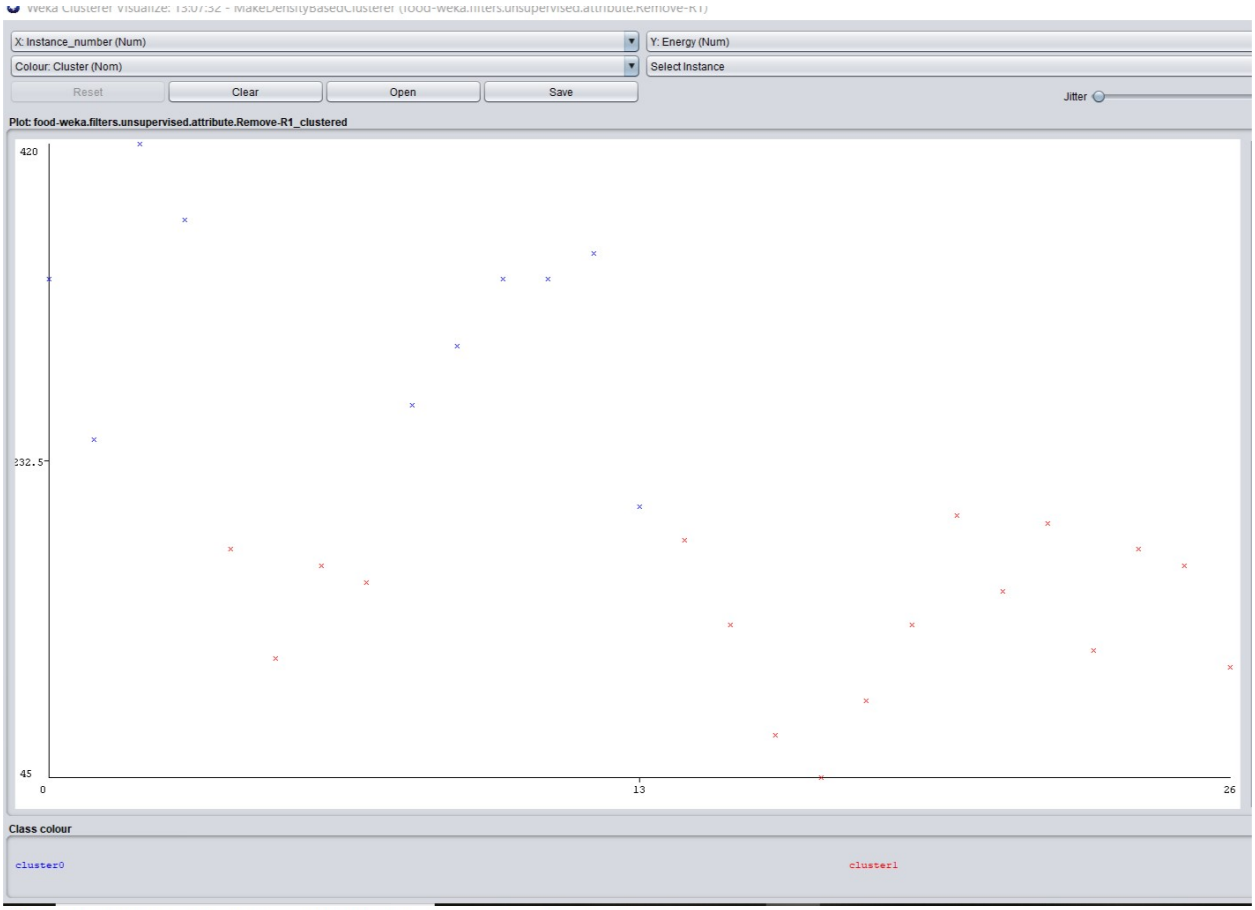


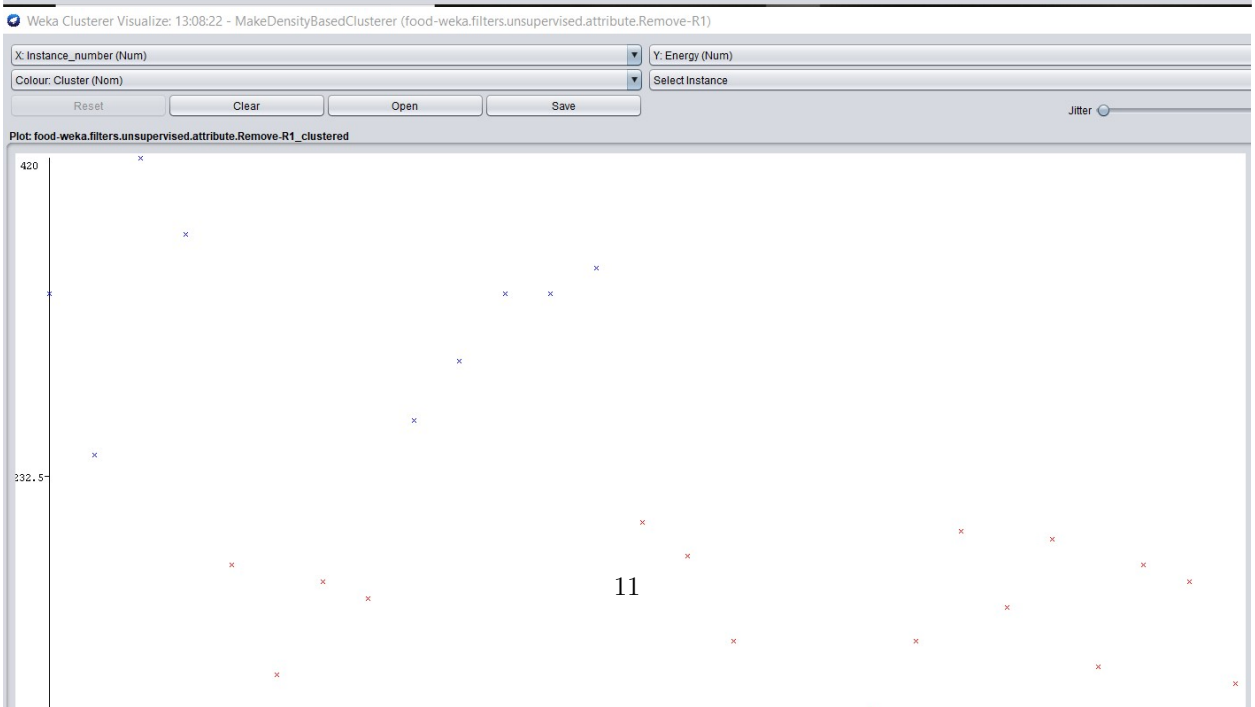
We chose k = 2 with seed 10. As can be seen in the plots above, the clearest separation is in terms of fat content. In particular, we can see that very fatty foods contain very little calcium. The blue cluster contains meats that are high in fat and contain very little calcium, while the red cluster contains meats that are low in fat.

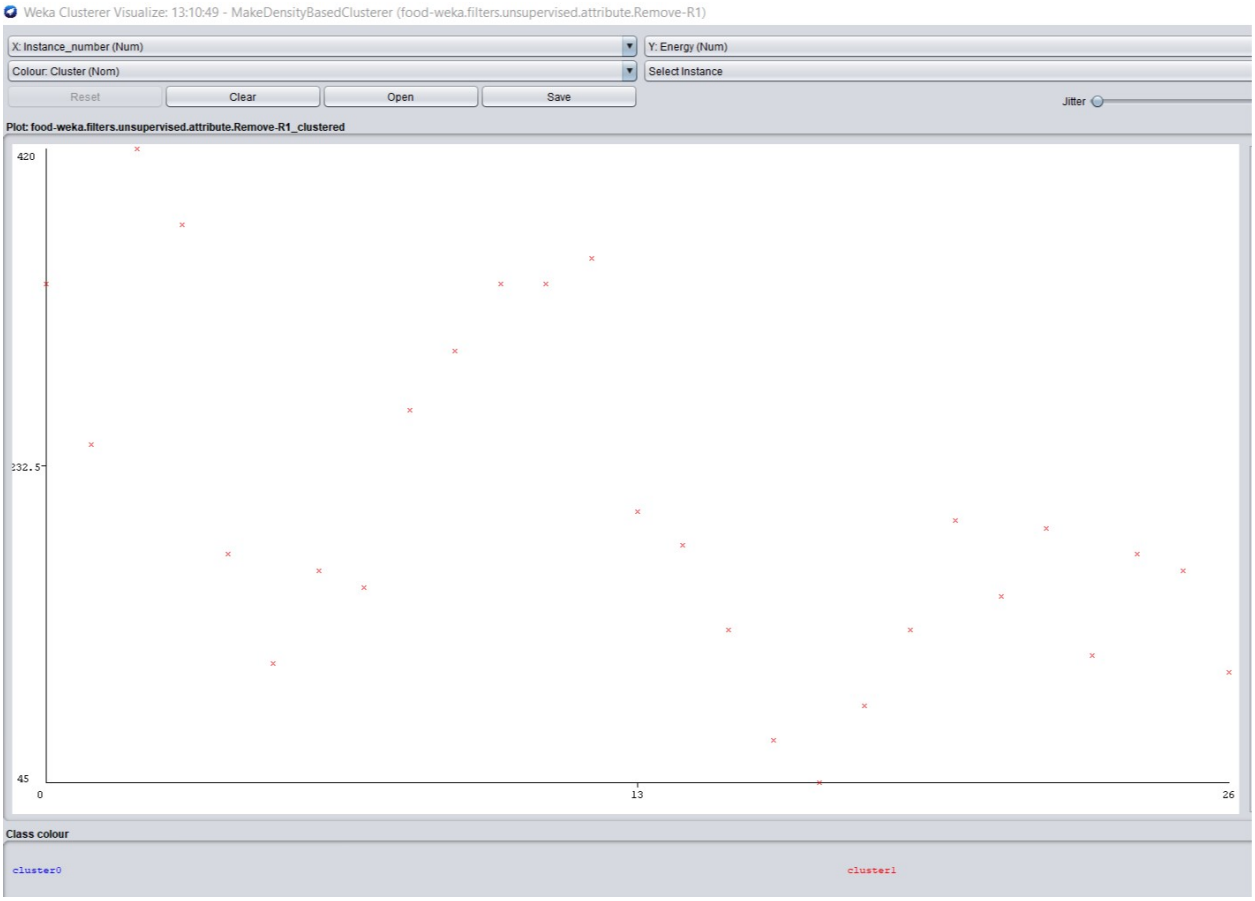
MakeDensityBasedClusters:

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)







In this method the algorithm creates clusters using K-means and then adjusts them with minimum standard deviation. The standard deviation value influencs the size of the final cluster. For a very small min standard deviation value we are guaranteed to have as many clusters as we asked for(ie. equal to the value of k we specified, eg. k=4). But for larger values of min standard deviation it starts merging the smaller clusters to satisfy the min standard deviation threshold, and so we end up with fewer number of clusters. So this also acts like a hyperparameter we use to control the cluster assignment, and the number of clusters.

With k = 2 we first used the default minimum of sd = 0.000001, and we ended up with two clusters with the same data distribution as with the k-means algorithm. For clustering with sd = 10 we got two clusters again, but two data points shifted cluster assignment. With sd = 1000 all elements end up being in one cluster as a result of no data points being outside this constraint on the minimum standard deviation.