Data Mining Lab-1

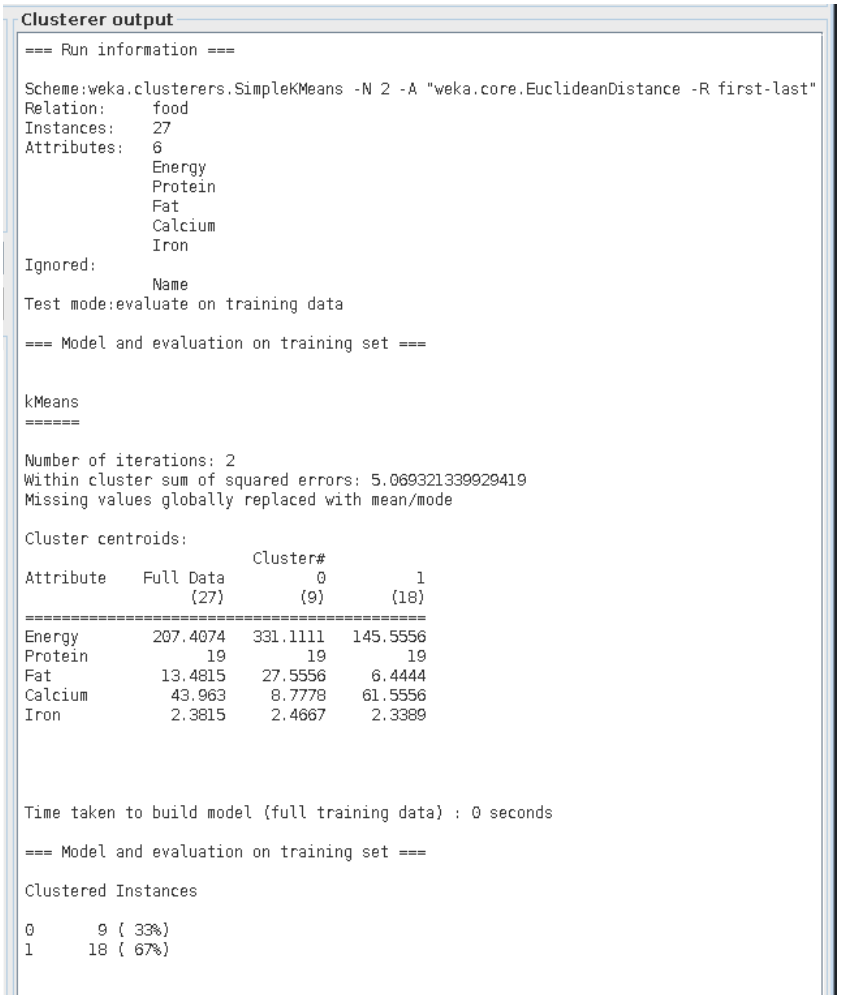
Names:

1. Siddhesh Sreedar (sidsr770)
2. Marijn Jaarsma (marja987)

Simple K-means

Q.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?)

We attempted clustering with k=2 and ignored the “Name” attribute based on intuition that will not provide any kind of valuable information in creating clusters (each name is unique).
As we can see below, the within-cluster SSE: 5.069

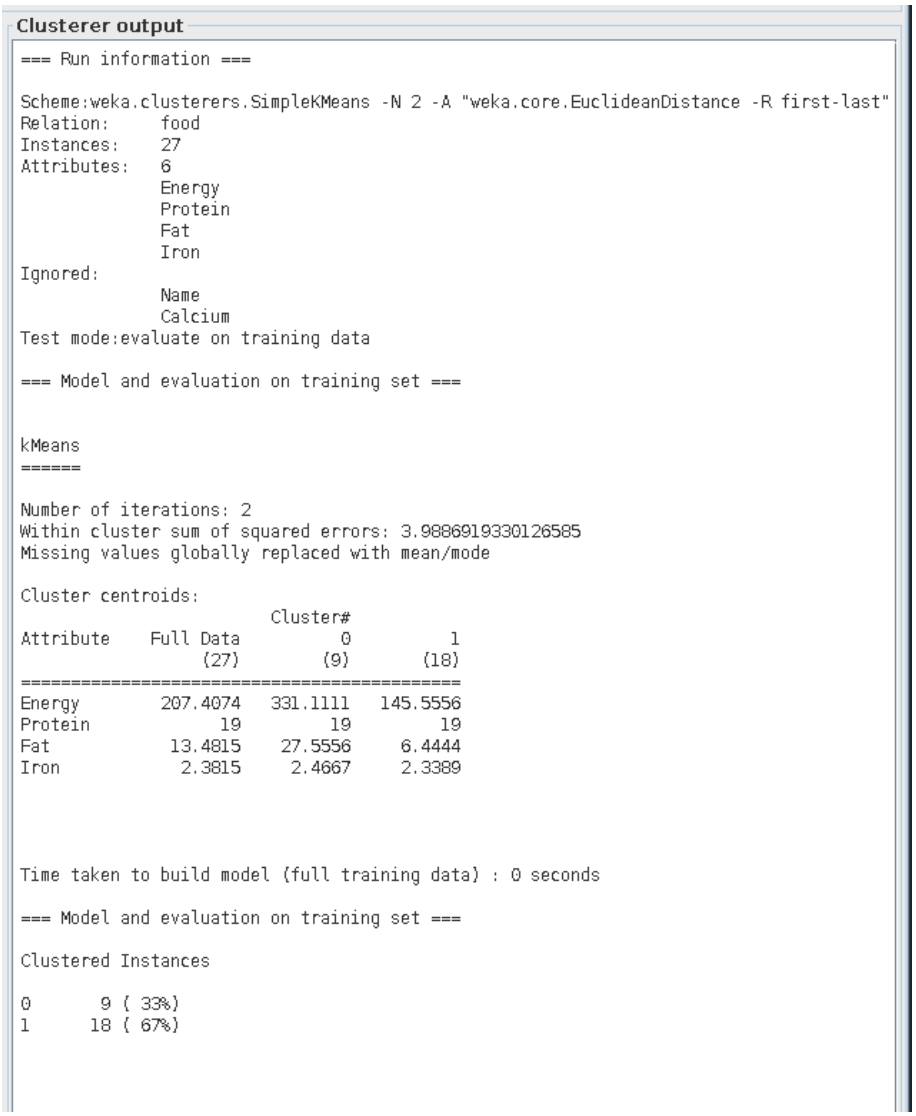


Clusterer output

```
=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last"
Relation:     food
Instances:    27
Attributes:   6
              Energy
              Protein
              Fat
              Calcium
              Iron
Ignored:
              Name
Test mode:evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 5.069321339929419
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1
                  (27)        (9)       (18)
=============================================
Energy         207.4074   331.1111   145.5556
Protein              19         19         19
Fat             13.4815    27.5556     6.4444
Calcium          43.963     8.7778    61.5556
Iron             2.3815     2.4667     2.3389



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       9 ( 33%)
1      18 ( 67%)
```

We noticed “calcium” has a large range of values including some outliers. Due to this, we are choosing to ignore this variable in the next iterations.

Second attempt clustering with k=2. We can see from below that the within-cluster SSE: 3.989

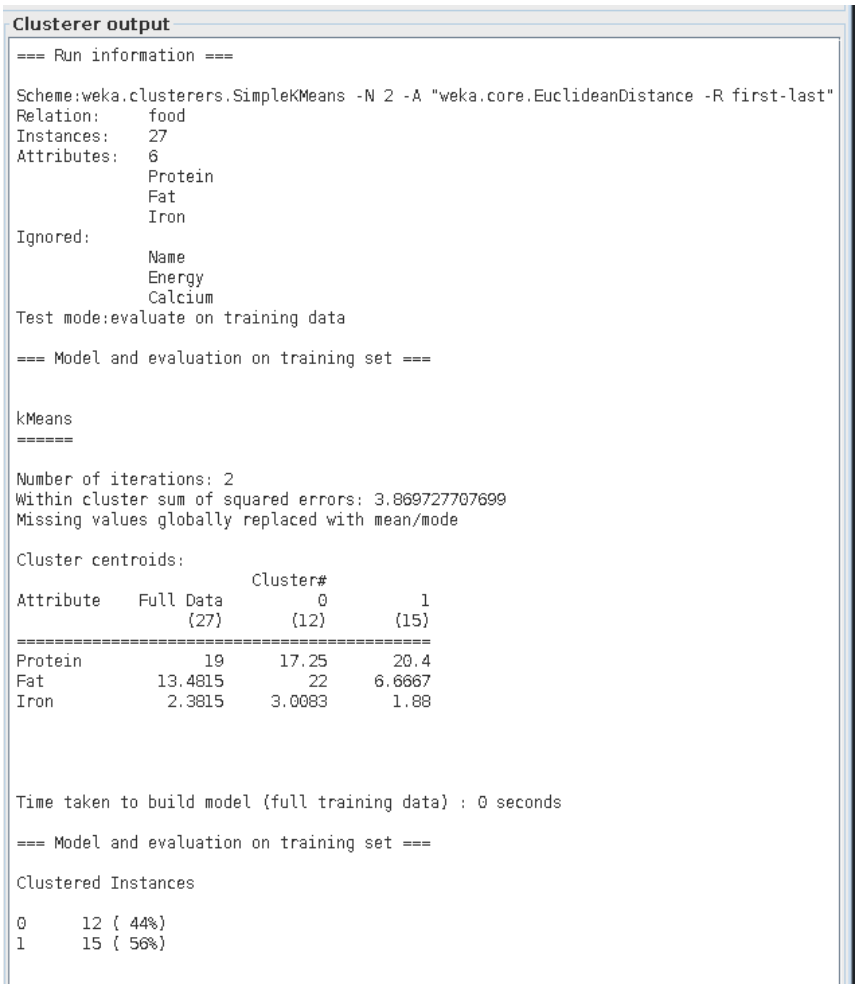


Clusterer output

```
=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last"
Relation:     food
Instances:    27
Attributes:   6
              Energy
              Protein
              Fat
              Iron
Ignored:
              Name
              Calcium
Test mode:evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 3.9886919330126585
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
                  (27)        (9)       (18)
=============================================
Energy        207.4074   331.1111   145.5556
Protein             19         19         19
Fat            13.4815    27.5556     6.4444
Iron            2.3815     2.4667     2.3389



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       9 ( 33%)
1      18 ( 67%)
```

We also removed the “energy” variable, as we feel that the “protein” and “fat” variables are likely correlated with this and it would not be fair to represent this multiple times. Also, the “energy” variable had a large range in values, similar to “calcium”. Additionally, as we were not sure whether the algorithm in Weka first normalizes the data we did not want this to unfairly skew the clusters towards energy.

In the third attempt clustering with k=2 and variables “protein”, “fat”, and “iron”. We can see from below that the within-cluster SSE: 3.870



Clusterer output

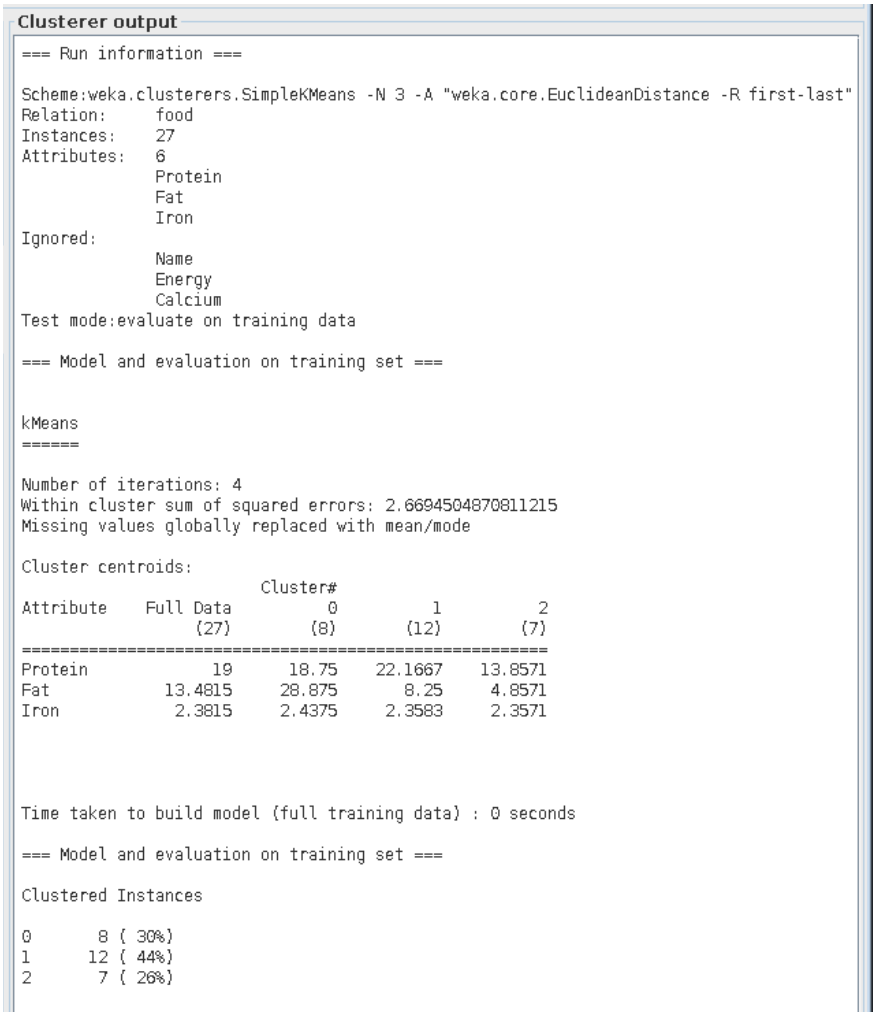
```
=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last"
Relation:     food
Instances:    27
Attributes:   6
              Protein
              Fat
              Iron
Ignored:
              Name
              Energy
              Calcium
Test mode:evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 3.869727707699
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1
                  (27)       (12)       (15)
==========================================
Protein             19      17.25       20.4
Fat            13.4815         22     6.6667
Iron            2.3815     3.0083       1.88



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      12 ( 44%)
1      15 ( 56%)
```

From the above information, we can see that the Foods in cluster 0 are slightly lower in protein, much higher in fat, and higher in iron. Additionally, we can see from the visualization that they are more energy-dense.

So we think that selecting only the “protein”, “fat”, and “iron” attributes is good enough.

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

We are now trying to increase the K value to 3 with the same variables as above i.e. “protein”, “fat”, and “iron”. We can see from below that the within-cluster SSE: 2.670



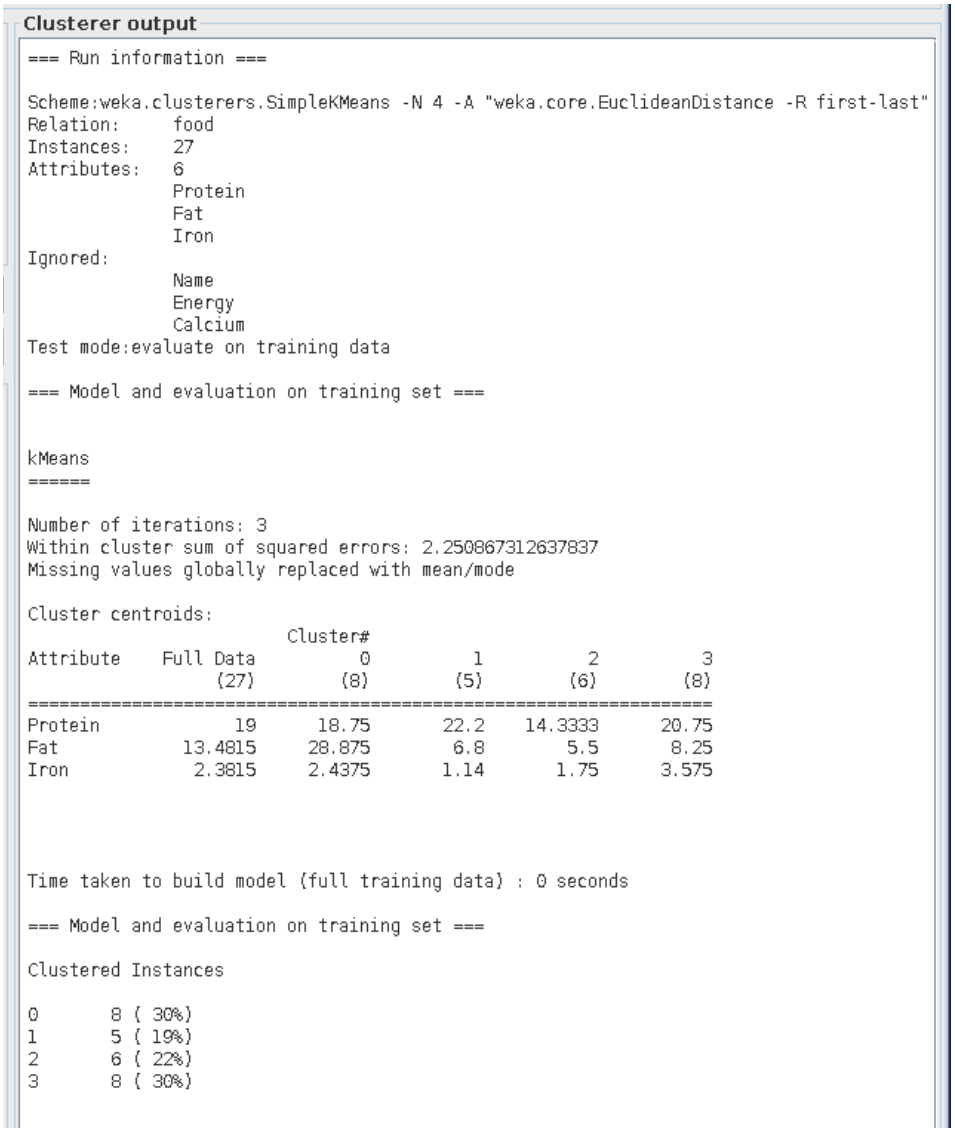
Clusterer output

```
=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 3 -A "weka.core.EuclideanDistance -R first-last"
Relation:     food
Instances:    27
Attributes:   6
              Protein
              Fat
              Iron
Ignored:
              Name
              Energy
              Calcium
Test mode:evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 2.6694504870811215
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2
                  (27)        (8)       (12)        (7)
=========================================================
Protein             19      18.75    22.1667    13.8571
Fat            13.4815     28.875       8.25     4.8571
Iron            2.3815     2.4375     2.3583     2.3571



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       8 ( 30%)
1      12 ( 44%)
2       7 ( 26%)
```

-

From the above information and from the visualizations, we can see that Cluster 0 has medium protein, very high fat, and is very energy-dense. Cluster 1 has the highest protein, low, but higher fat than cluster 2, and has mid energy. Cluster 2 has low protein, low fat, and low energy as well as high calcium.

We run the algorithm again with K = 4 and with the same variables as above just to gain more insights. We can see from below that the within-cluster SSE: 2.251



Clusterer output

```
=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 4 -A "weka.core.EuclideanDistance -R first-last"
Relation:     food
Instances:    27
Attributes:   6
              Protein
              Fat
              Iron
Ignored:
              Name
              Energy
              Calcium
Test mode:evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 2.250867312637837
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3
                  (27)        (8)        (5)        (6)        (8)
=============================================================================
Protein             19      18.75       22.2    14.3333      20.75
Fat            13.4815     28.875        6.8        5.5       8.25
Iron            2.3815     2.4375       1.14       1.75      3.575



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       8 ( 30%)
1       5 ( 19%)
2       6 ( 22%)
3       8 ( 30%)
```

From the above information and the visualizations, we can see that Cluster 0 clearly contains high-energy foods, while the rest of the clusters are mixed around the mid to low range. Cluster 2 has lower protein, while the rest are mixed. Cluster 0 has high fat, the rest are mixed. Clusters 2 and 3 have higher calcium. Clusters 2 and 3 have a few high outliers for iron, but the rest is mixed.

With this clustering setup, it seems like we're overdoing it. Some of the cluster centroids are very similar, perhaps picking up data within the same cluster. Additionally, the cluster sizes are becoming very small here. **Therefore we think that 3 clusters seem like the ideal setup.**

Q.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.

Using a different seed (78367383) we can see that it produces similar results in terms of cluster centroids as the previous seed value i.e 10, but the number of instances per cluster is slightly different as well as the within-cluster SSE (1.934) as we can see below:

Clusterer output

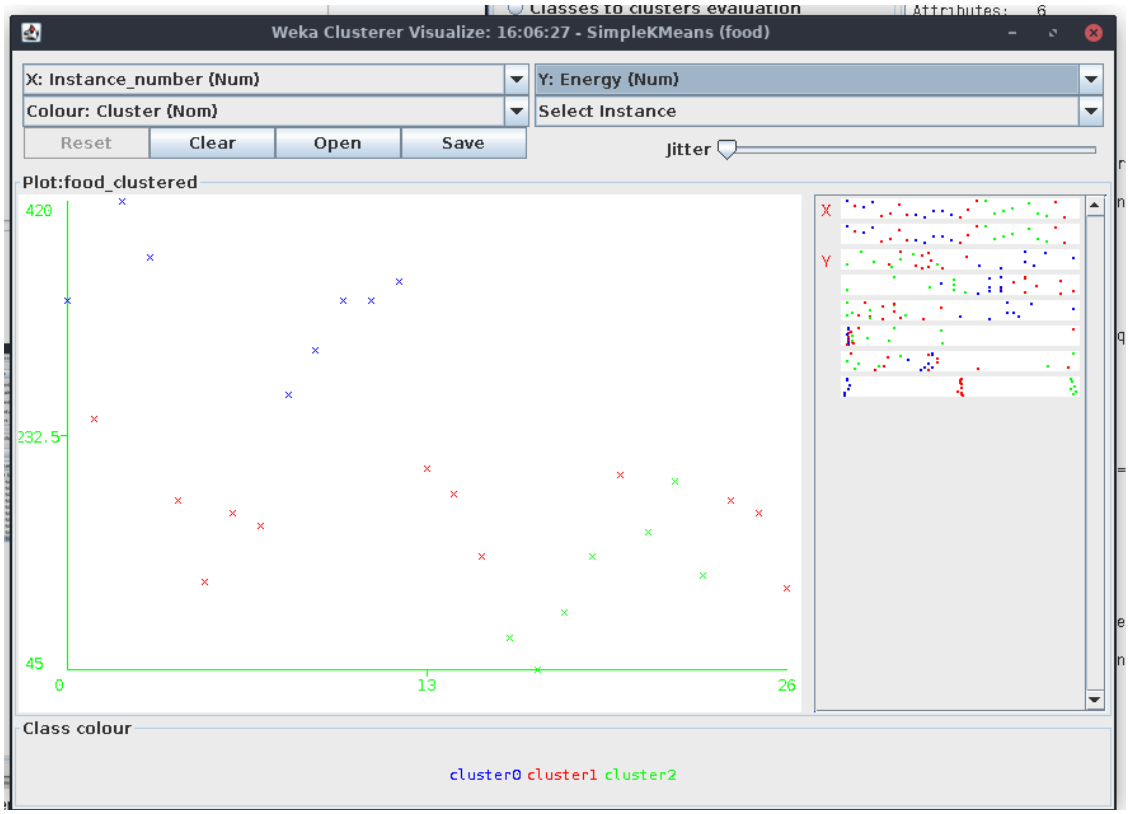
```
=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 3 -A "weka.core.EuclideanDistance -R first-last"
Relation:     food
Instances:    27
Attributes:   6
              Protein
              Fat
              Iron
Ignored:
              Name
              Energy
              Calcium
Test mode:evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 1.9337202960097069
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2
                (27)        (8)       (16)        (3)
=========================================================
Protein             19      18.75    19.9375    14.6667
Fat            13.4815     28.875     7.875     2.3333
Iron            2.3815     2.4375     1.7188     5.7667



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       8 ( 30%)
1      16 ( 59%)
2       3 ( 11%)
```

The seed value controls the initial random setup of the centroids. If the K-means algorithm starts with the centroids in a different spot, it might converge to a different result.

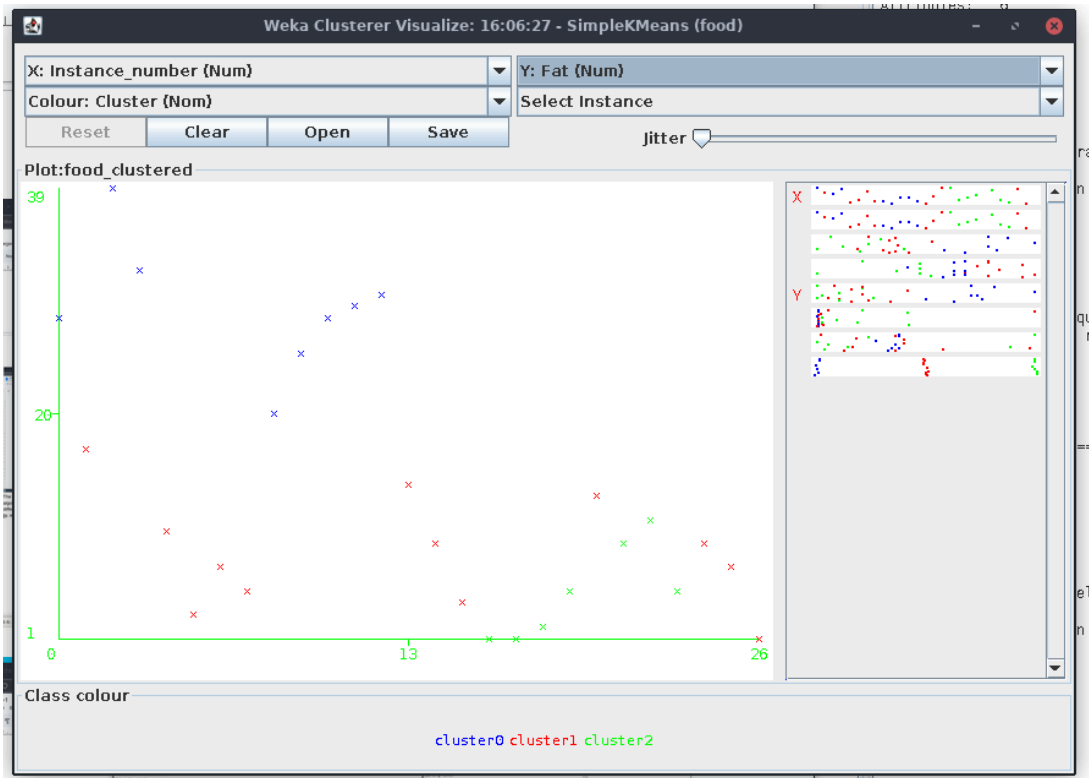
Q.4) Do you think the clusters are "good" clusters? (Are all of its members "similar" to each other? Are members from different clusters dissimilar?)

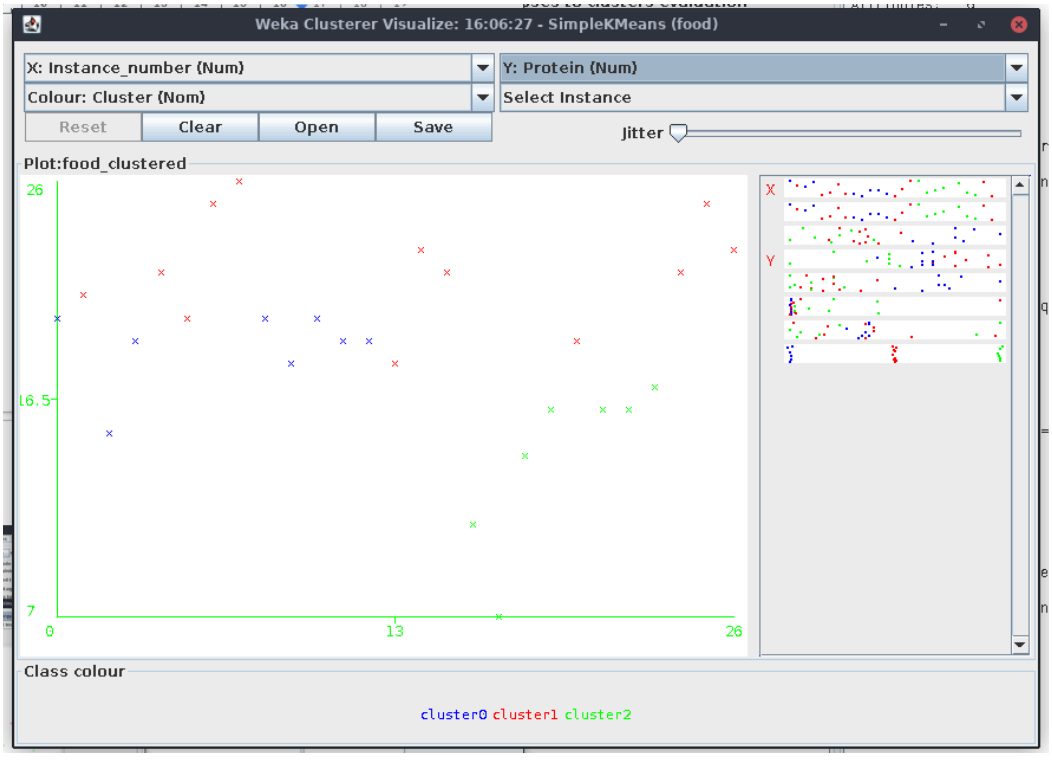
Upon analyzing the information outputted by the algorithm and the visualizations provided, we do think that the 3 clusters are good clusters.

Based on the below plot for “energy”, we see a clear distinction between the three clusters.



And also for the two plots for “fat” and “protein” below respectively, we can see a good cluster formation but with a few points being similar in different clusters.


Weka Clusterer Visualize: 16:06:27 - SimpleKMeans (food)
X: Instance_number (Num)
Y: Fat (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:food_clustered
39
20
1
0
13
26
X
Y
Class colour
cluster0 cluster1 cluster2


Weka Clusterer Visualize: 16:06:27 - SimpleKMeans (food)
X: Instance_number (Num)
Y: Protein (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:food_clustered
26
16.5
7
0
13
26
X
Y
Class colour
cluster0 cluster1 cluster2

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

Cluster 0: High Energy
Cluster 1: Mid Energy
Cluster 2: Low Energy

MakeDensityBasedClusterer

Q.1) Use the SimpleKMeans clusterer which gave the result you haven chosen in 5).

```
Number of iterations: 4
Within cluster sum of squared errors: 2.6694504870811215
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute    Full Data          0          1          2
                  (27)        (8)       (12)        (7)
=========================================================
Protein             19      18.75    22.1667    13.8571
Fat            13.4815     28.875       8.25     4.8571
Iron            2.3815     2.4375     2.3583     2.3571


Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.3

Attribute: Protein
Normal Distribution. Mean = 18.75 StdDev = 1.5612
Attribute: Fat
Normal Distribution. Mean = 28.875 StdDev = 5.1097
Attribute: Iron
Normal Distribution. Mean = 2.4375 StdDev = 0.1932

Cluster: 1 Prior probability: 0.4333

Attribute: Protein
Normal Distribution. Mean = 22.1667 StdDev = 2.3393
Attribute: Fat
Normal Distribution. Mean = 8.25 StdDev = 4.5484
Attribute: Iron
Normal Distribution. Mean = 2.3583 StdDev = 1.3726

Cluster: 2 Prior probability: 0.2667

Attribute: Protein
Normal Distribution. Mean = 13.8571 StdDev = 3.3564
Attribute: Fat
Normal Distribution. Mean = 4.8571 StdDev = 3.6422
Attribute: Iron
Normal Distribution. Mean = 2.3571 StdDev = 2.1573


Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       8 ( 30%)
1      12 ( 44%)
2       7 ( 26%)


Log likelihood: -7.41819
```

Q.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)

We started with the initial value for minimum standard deviation, 1e-6, and got the same clusters as we got from the simple kmeans. Then we changed the minimum sd to 10 and a lot of the data points went into cluster 1 to achieve this. When setting the minimum sd to an extreme value such as 100, all the values are grouped into one cluster.

```
Number of iterations: 4
Within cluster sum of squared errors: 2.6694504870811215
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2
                  (27)        (8)       (12)        (7)
=========================================================
Protein             19      18.75    22.1667    13.8571
Fat            13.4815     28.875       8.25     4.8571
Iron            2.3815     2.4375     2.3583     2.3571


Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.3

Attribute: Protein
Normal Distribution. Mean = 18.75 StdDev = 10
Attribute: Fat
Normal Distribution. Mean = 28.875 StdDev = 11.257
Attribute: Iron
Normal Distribution. Mean = 2.4375 StdDev = 10

Cluster: 1 Prior probability: 0.4333

Attribute: Protein
Normal Distribution. Mean = 22.1667 StdDev = 10
Attribute: Fat
Normal Distribution. Mean = 8.25 StdDev = 11.257
Attribute: Iron
Normal Distribution. Mean = 2.3583 StdDev = 10

Cluster: 2 Prior probability: 0.2667

Attribute: Protein
Normal Distribution. Mean = 13.8571 StdDev = 10
Attribute: Fat
Normal Distribution. Mean = 4.8571 StdDev = 11.257
Attribute: Iron
Normal Distribution. Mean = 2.3571 StdDev = 10


Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       7 ( 26%)
1      18 ( 67%)
2       2 (  7%)


Log likelihood: -10.45387
```

```
Number of iterations: 4
Within cluster sum of squared errors: 2.6694504870811215
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2
                  (27)        (8)       (12)        (7)
=========================================================
Protein             19      18.75    22.1667    13.8571
Fat            13.4815     28.875       8.25     4.8571
Iron            2.3815     2.4375     2.3583     2.3571


Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.3

Attribute: Protein
Normal Distribution. Mean = 18.75 StdDev = 100
Attribute: Fat
Normal Distribution. Mean = 28.875 StdDev = 100
Attribute: Iron
Normal Distribution. Mean = 2.4375 StdDev = 100

Cluster: 1 Prior probability: 0.4333

Attribute: Protein
Normal Distribution. Mean = 22.1667 StdDev = 100
Attribute: Fat
Normal Distribution. Mean = 8.25 StdDev = 100
Attribute: Iron
Normal Distribution. Mean = 2.3583 StdDev = 100

Cluster: 2 Prior probability: 0.2667

Attribute: Protein
Normal Distribution. Mean = 13.8571 StdDev = 100
Attribute: Fat
Normal Distribution. Mean = 4.8571 StdDev = 100
Attribute: Iron
Normal Distribution. Mean = 2.3571 StdDev = 100


Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

1      27 (100%)


Log likelihood: -16.58504
```