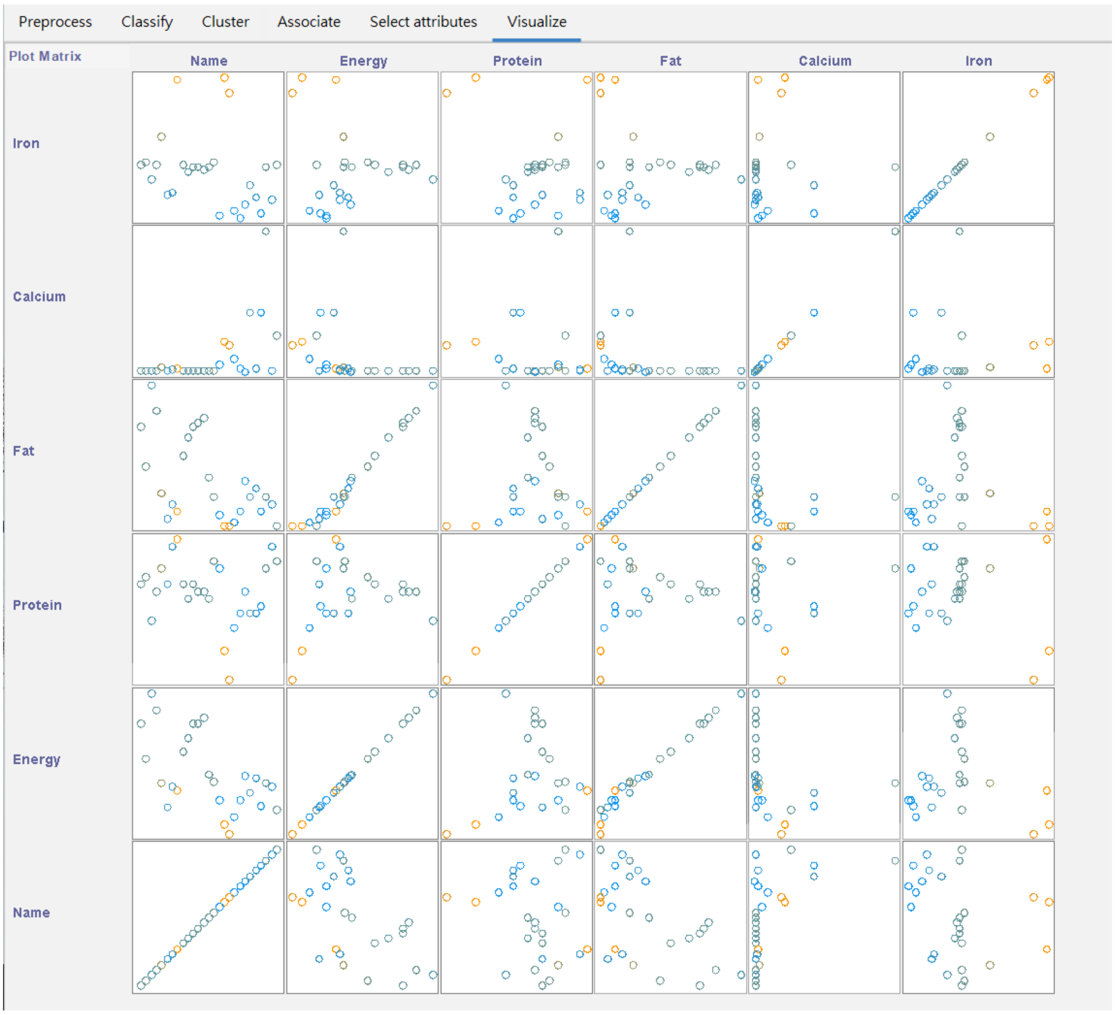
Q: 1. Choose a set of attributes for clustering and give a motivation. (Hint: always ignore attribute "name". Why does the name attribute need to be ignored?)*

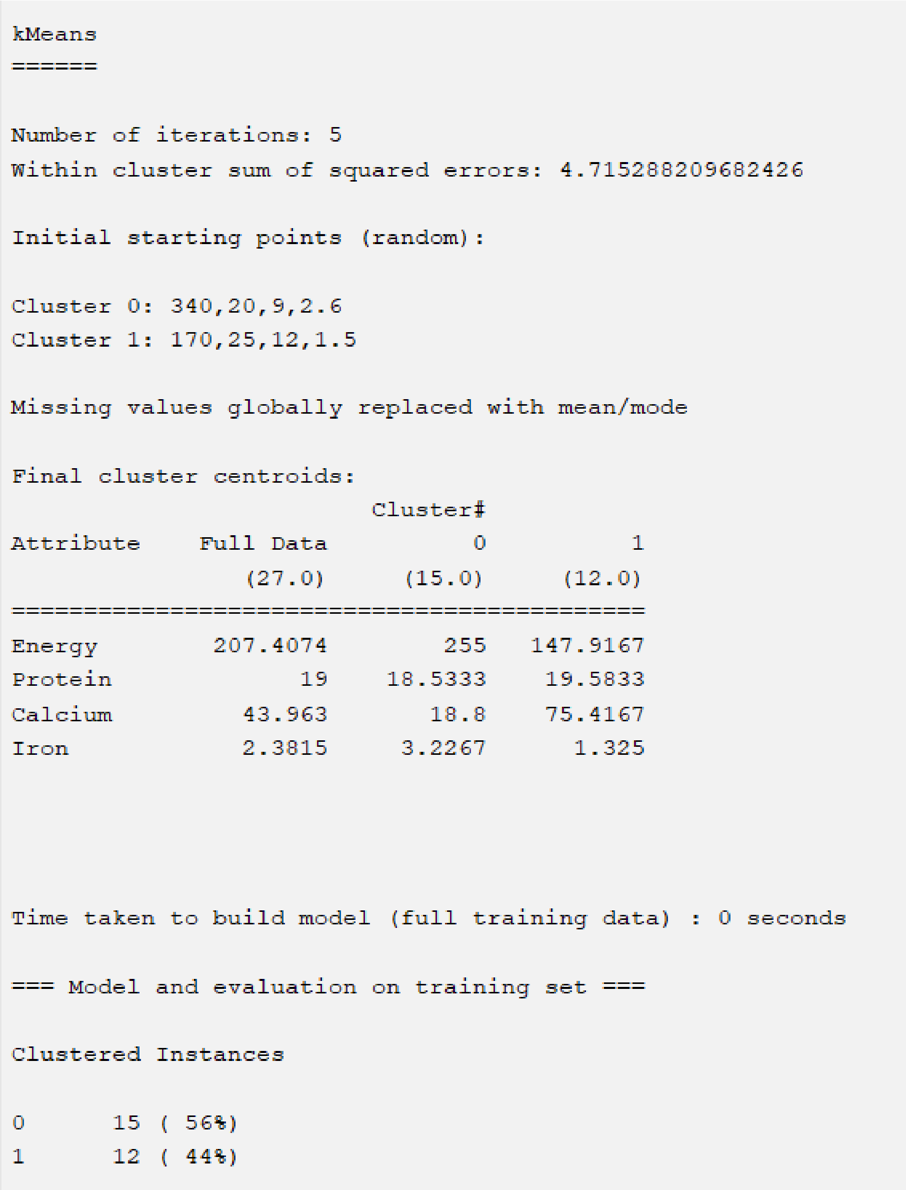


To choose the attributes we want to use when clustering. We first observed all the attributes' correlation to each other. As the above graph shows, we observed that fat and energy have a very high correlation. Therefore we can choose to ignore one of them, and we end up ignoring fat in this assignment.

We always ignore the attribute name because it is a categorical label and will not contribute to the clustering process in any way.

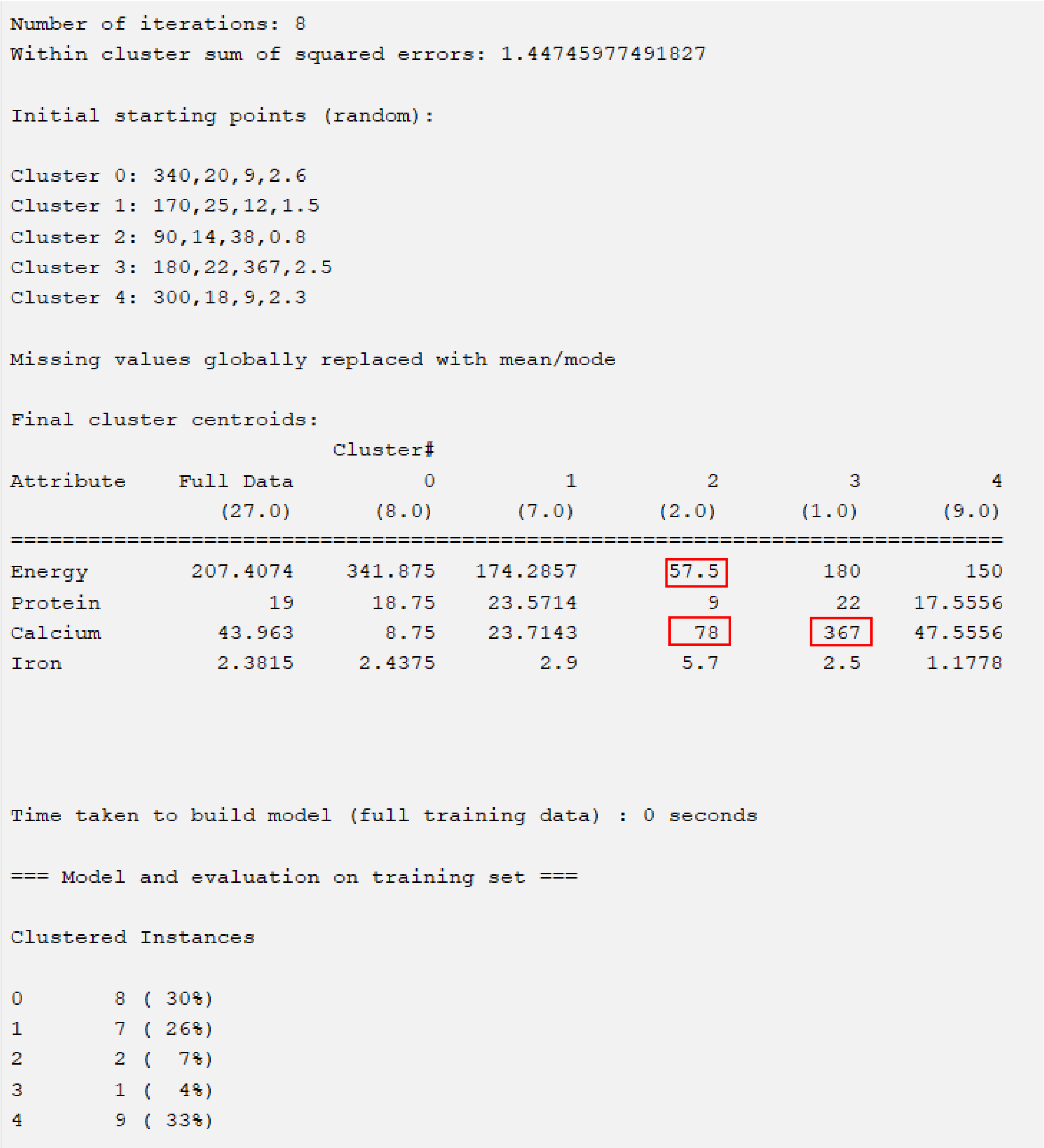
2. Experiment with at least two different numbers of clusters, e.g. 2 and 5, but with the same seed value 10.

We have try two different numbers of clusters 2 and 5. The result are as below.



```
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 4.715288209682426

Initial starting points (random):

Cluster 0: 340,20,9,2.6
Cluster 1: 170,25,12,1.5

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
                (27.0)     (15.0)     (12.0)
============================================
Energy        207.4074        255   147.9167
Protein             19    18.5333    19.5833
Calcium         43.963       18.8    75.4167
Iron            2.3815     3.2267      1.325




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

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

Clustered Instances

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

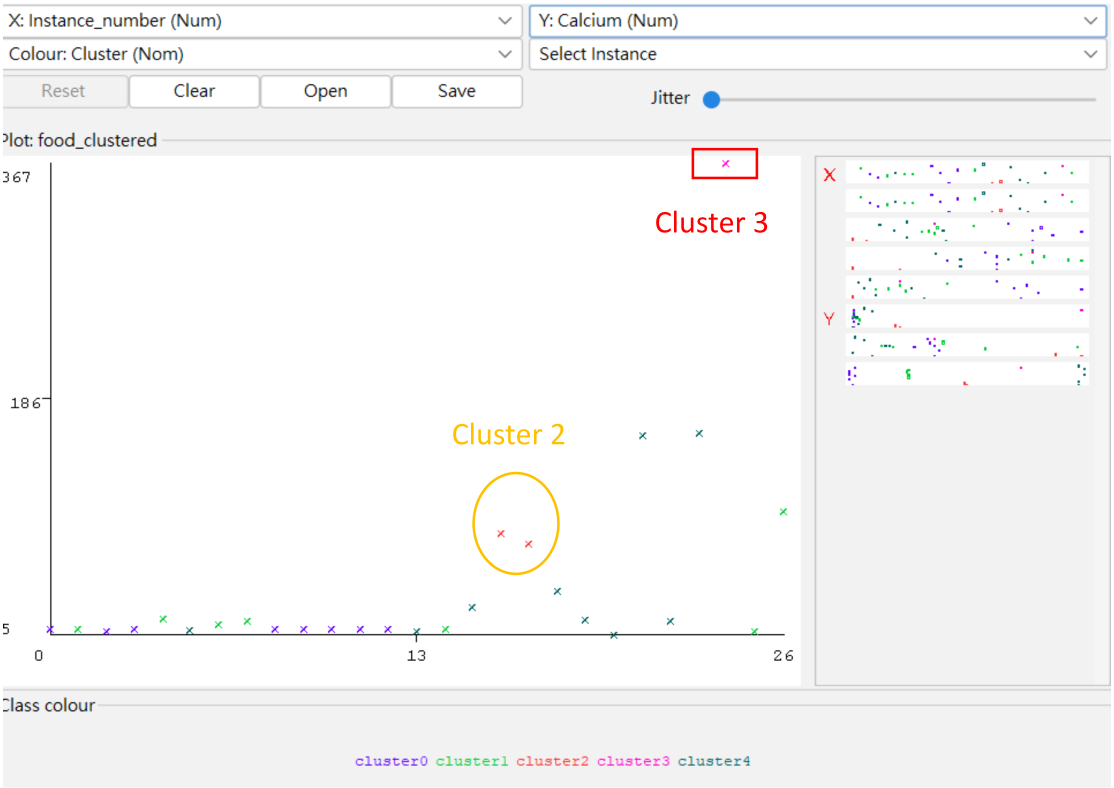
2 Clusters



```
Number of iterations: 8
Within cluster sum of squared errors: 1.44745977491827

Initial starting points (random):

Cluster 0: 340,20,9,2.6
Cluster 1: 170,25,12,1.5
Cluster 2: 90,14,38,0.8
Cluster 3: 180,22,367,2.5
Cluster 4: 300,18,9,2.3

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1          2          3          4
                (27.0)      (8.0)      (7.0)      (2.0)      (1.0)      (9.0)
=============================================================================
Energy        207.4074    341.875   174.2857       57.5        180        150
Protein             19      18.75    23.5714          9         22    17.5556
Calcium         43.963       8.75    23.7143         78        367    47.5556
Iron            2.3815     2.4375        2.9        5.7        2.5     1.1778




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

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

Clustered Instances

0       8 ( 30%)
1       7 ( 26%)
2       2 (  7%)
3       1 (  4%)
4       9 ( 33%)
```

5 Clusters

As we can see from the above result, we can see that in the 5-clusters case, two clusters have only one and two objects. It is because there is one outliers in Calcium that has higher value and two object with the combination of lower engery and higher than average Calicum, as the below graph shown



X: Instance_number (Num)

Y: Calcium (Num)

Colour: Cluster (Nom)

Select Instance

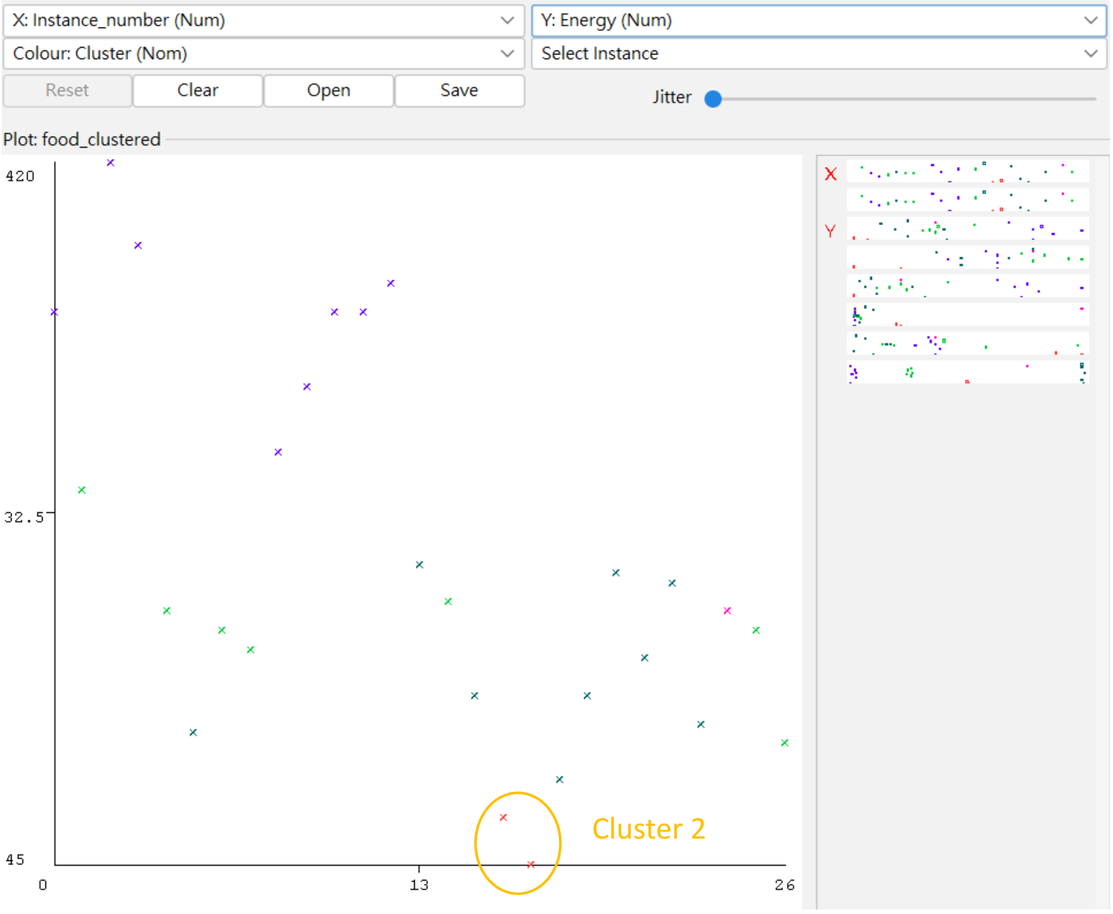
Plot: food_clustered

Class colour: cluster0 cluster1 cluster2 cluster3 cluster4

Cluster 3

Cluster 2

Outlier in Calcium



X: Instance_number (Num)

Y: Energy (Num)

Colour: Cluster (Nom)

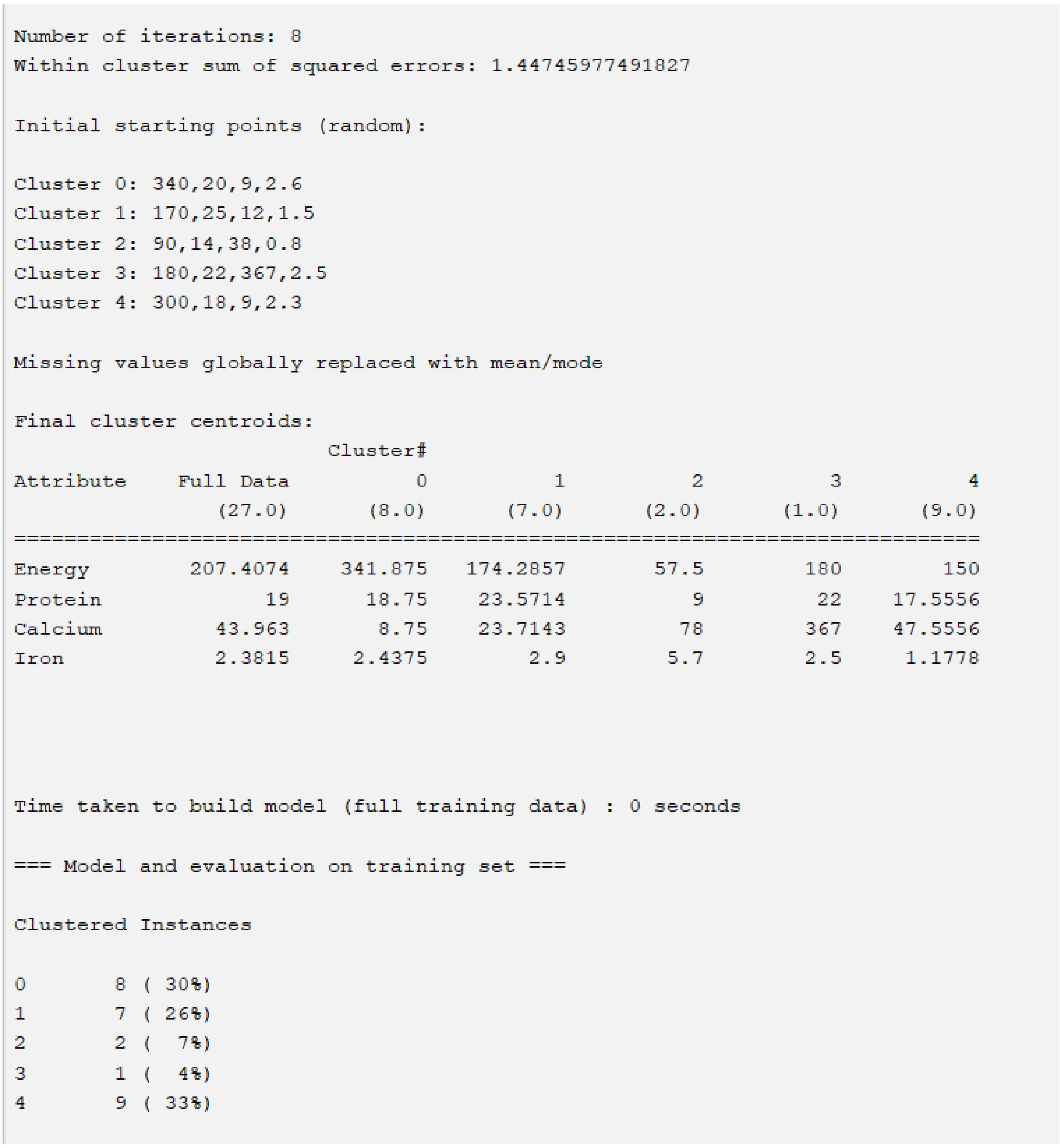
Select Instance

Plot: food_clustered

Cluster 2

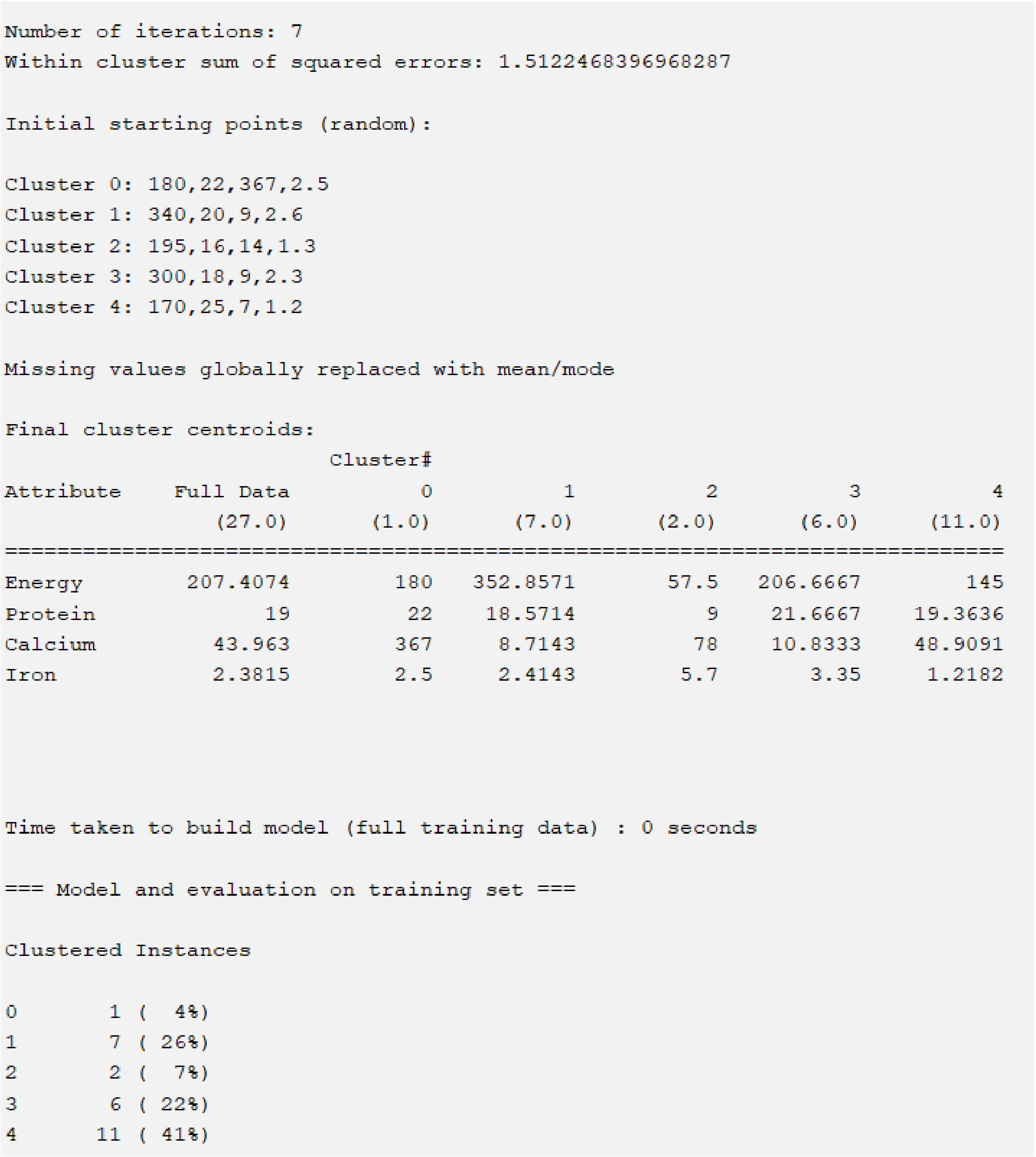
3. Then try with a different seed value, i.e. different initial cluster centers. Compare the results with the previous results. Explain what the seed value controls.

We compare seed 10 and seed 1 for 5 clusters cases, and the result are as below



```
Number of iterations: 8
Within cluster sum of squared errors: 1.44745977491827

Initial starting points (random):

Cluster 0: 340,20,9,2.6
Cluster 1: 170,25,12,1.5
Cluster 2: 90,14,38,0.8
Cluster 3: 180,22,367,2.5
Cluster 4: 300,18,9,2.3

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute     Full Data          0          1          2          3          4
                 (27.0)      (8.0)      (7.0)      (2.0)      (1.0)      (9.0)
=============================================================================
Energy         207.4074    341.875   174.2857       57.5        180        150
Protein              19      18.75    23.5714          9         22    17.5556
Calcium          43.963       8.75    23.7143         78        367    47.5556
Iron             2.3815     2.4375        2.9        5.7        2.5     1.1778




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

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

Clustered Instances

0       8 ( 30%)
1       7 ( 26%)
2       2 (  7%)
3       1 (  4%)
4       9 ( 33%)
```

Seed=10



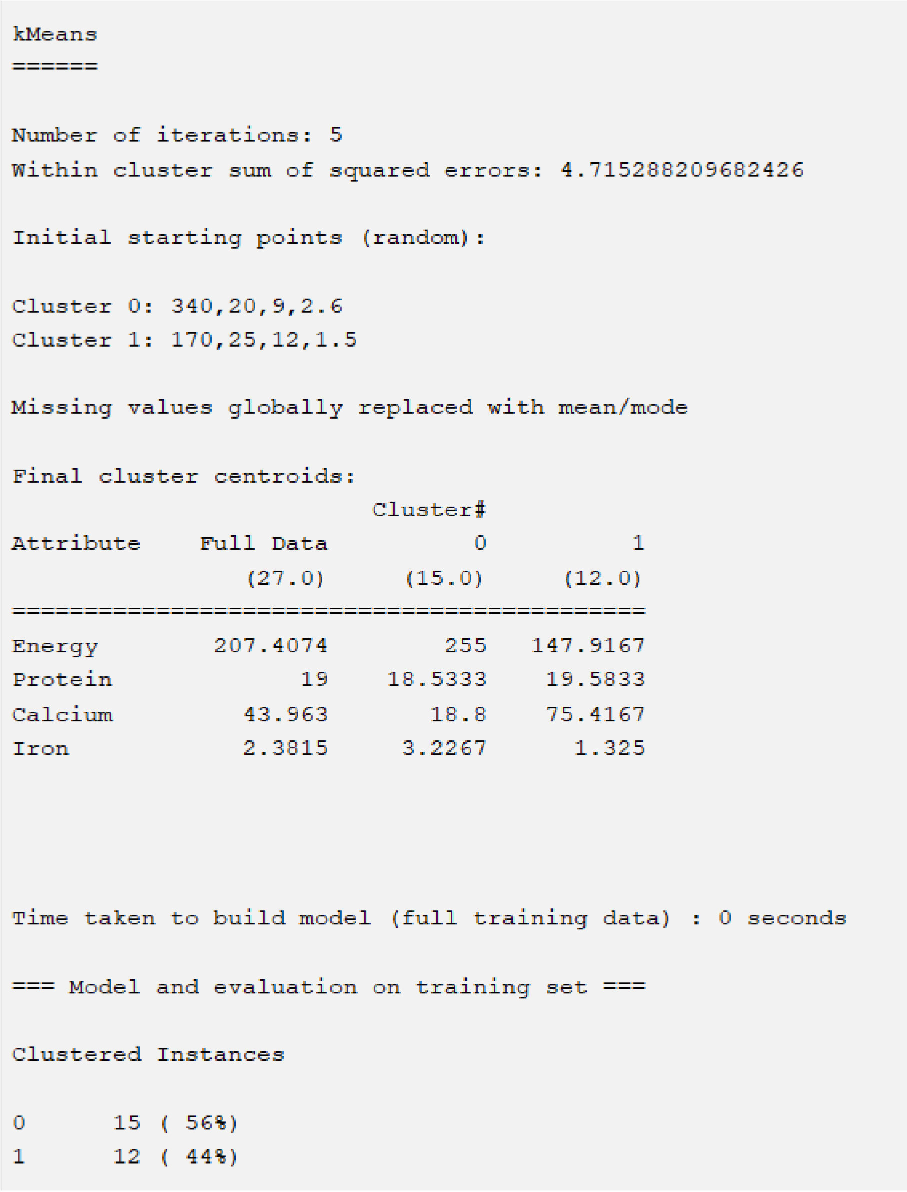
```
Number of iterations: 7
Within cluster sum of squared errors: 1.5122468396968287

Initial starting points (random):

Cluster 0: 180,22,367,2.5
Cluster 1: 340,20,9,2.6
Cluster 2: 195,16,14,1.3
Cluster 3: 300,18,9,2.3
Cluster 4: 170,25,7,1.2

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1          2          3          4
                (27.0)      (1.0)      (7.0)      (2.0)      (6.0)     (11.0)
==============================================================================
Energy        207.4074        180   352.8571       57.5   206.6667        145
Protein             19         22    18.5714          9    21.6667    19.3636
Calcium         43.963        367     8.7143         78    10.8333    48.9091
Iron            2.3815        2.5     2.4143        5.7       3.35     1.2182




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

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

Clustered Instances

0        1 (  4%)
1        7 ( 26%)
2        2 (  7%)
3        6 ( 22%)
4       11 ( 41%)
```

Seed=5

As it can be observed, the seed will effect the initial starting point of the clusters, and therefore affect some of the clusters. We can see that both cases kept the outliers we dicussed in the previous question, but changes the number of objects in other clusters.

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

Ans: For this and below question, we use k = 2 clusters.



```
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 4.715288209682426

Initial starting points (random):

Cluster 0: 340,20,9,2.6
Cluster 1: 170,25,12,1.5

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute     Full Data         0          1
                 (27.0)    (15.0)     (12.0)
============================================
Energy         207.4074       255   147.9167
Protein              19   18.5333    19.5833
Calcium          43.963      18.8    75.4167
Iron             2.3815    3.2267      1.325




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

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

Clustered Instances

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

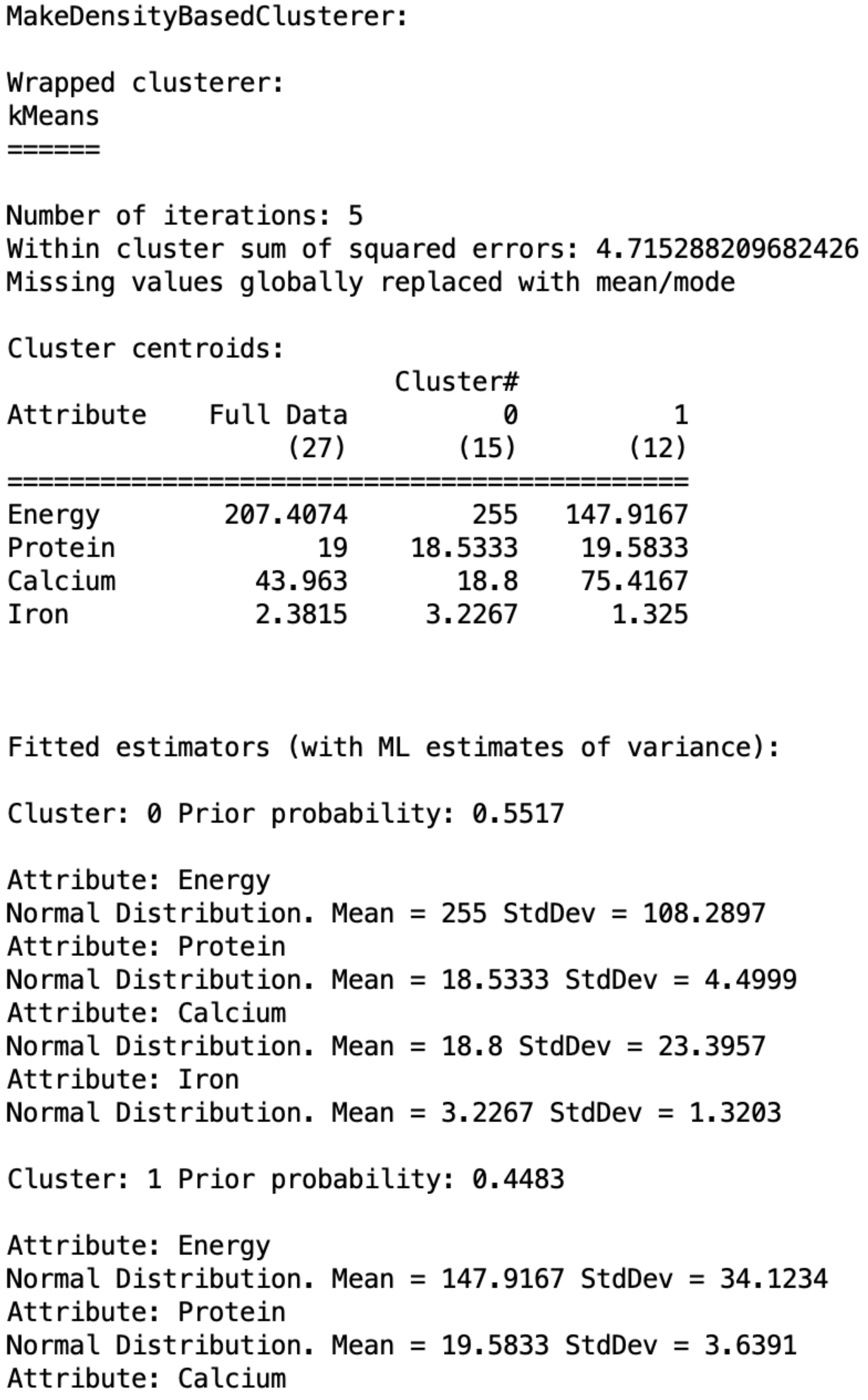
Each clusters represents a category of food.
Cluster #0 is characterized by high average energy, low average calcium,higher average iron food types (Nuts, cereals)

Cluster #1 is charcterized by low average energy, high average calcium,lower average iron food type (Leafy greens, tofu)

MakeDensityBasedClusters

Q. Experiment with at least two different standard deviations. Compare the results.

K = 2 with MinStdDev set to 1e-6

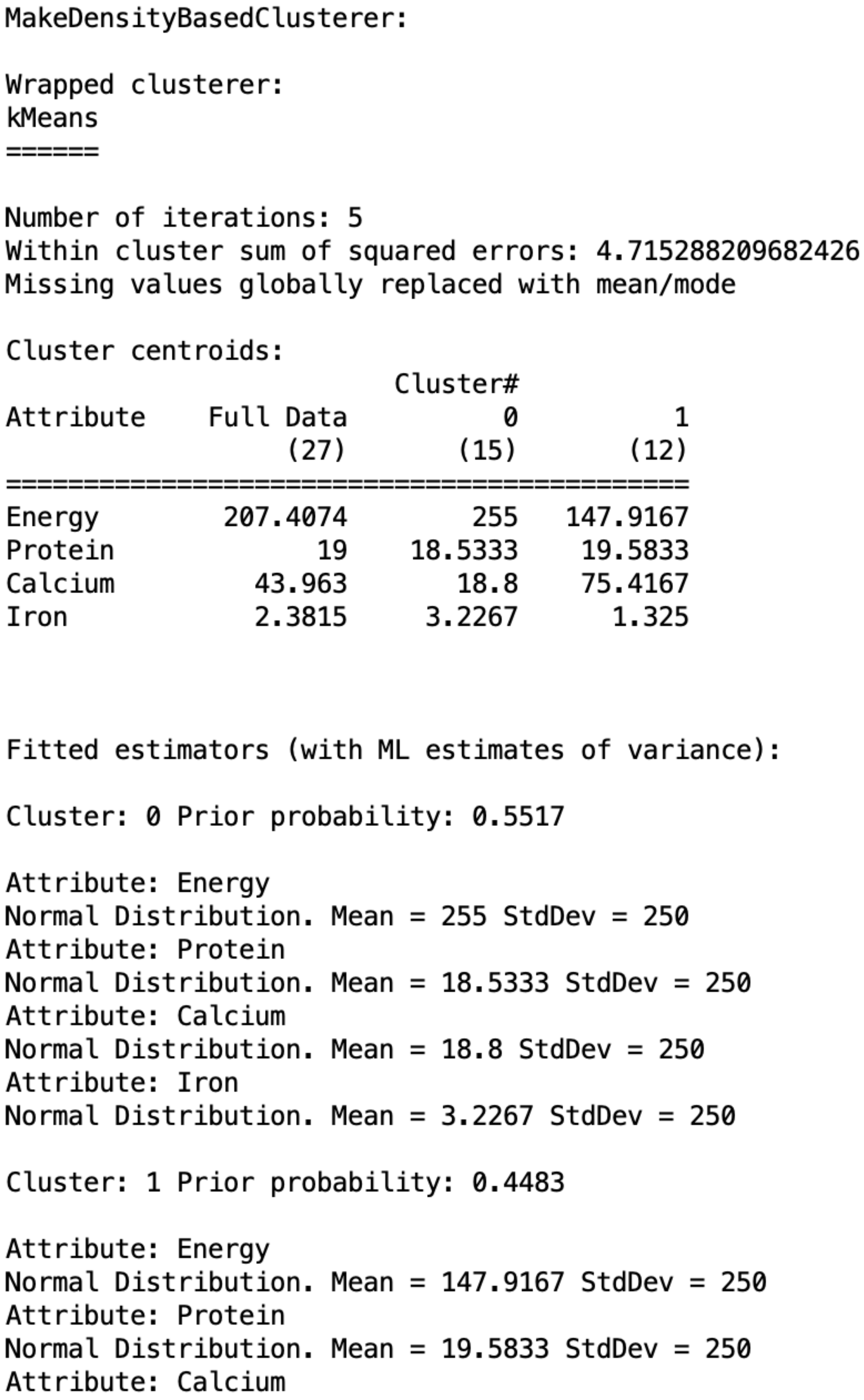


```
MakeDensityBasedClusterer:

Wrapped clusterer:
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 4.715288209682426
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
                  (27)       (15)       (12)
============================================
Energy        207.4074        255   147.9167
Protein             19    18.5333    19.5833
Calcium         43.963       18.8    75.4167
Iron            2.3815     3.2267      1.325



Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.5517

Attribute: Energy
Normal Distribution. Mean = 255 StdDev = 108.2897
Attribute: Protein
Normal Distribution. Mean = 18.5333 StdDev = 4.4999
Attribute: Calcium
Normal Distribution. Mean = 18.8 StdDev = 23.3957
Attribute: Iron
Normal Distribution. Mean = 3.2267 StdDev = 1.3203

Cluster: 1 Prior probability: 0.4483

Attribute: Energy
Normal Distribution. Mean = 147.9167 StdDev = 34.1234
Attribute: Protein
Normal Distribution. Mean = 19.5833 StdDev = 3.6391
Attribute: Calcium
```

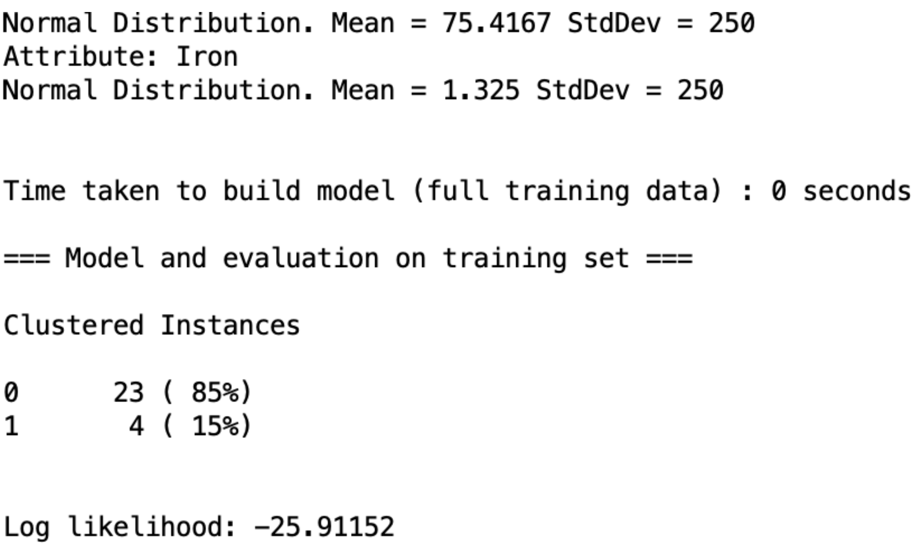
```
Normal Distribution. Mean = 75.4167 StdDev = 103.5788
Attribute: Iron
Normal Distribution. Mean = 1.325 StdDev = 0.6622


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

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

Clustered Instances

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


Log likelihood: -15.63591
```

K = 2 with minStdDev = 250

Results :

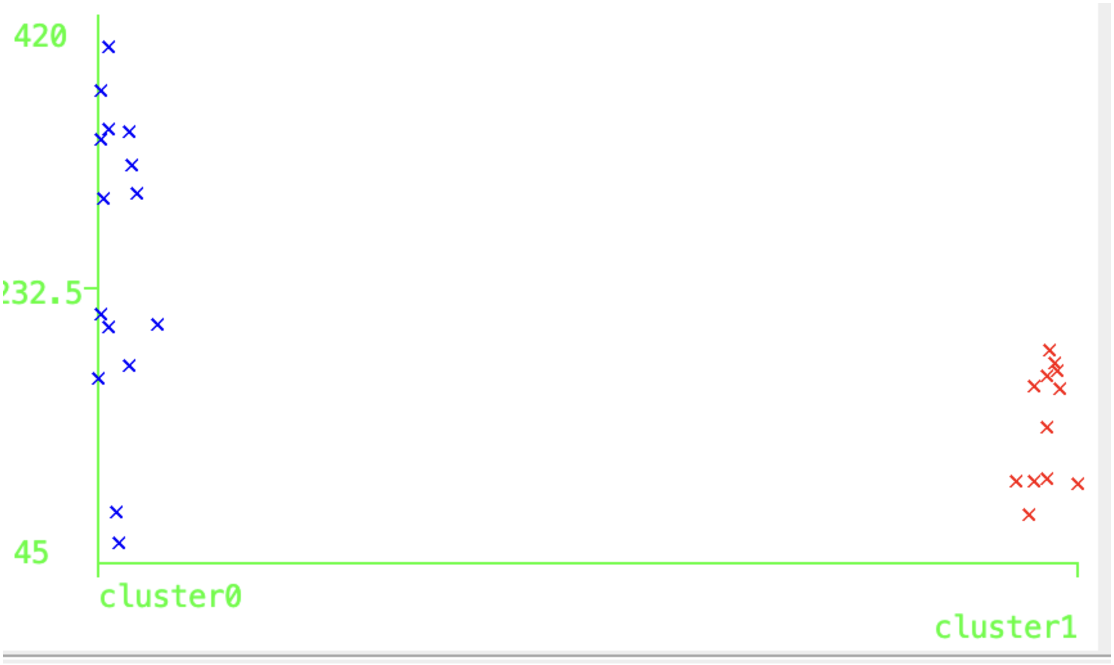


```
MakeDensityBasedClusterer:

Wrapped clusterer:
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 4.715288209682426
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
                  (27)       (15)       (12)
============================================
Energy        207.4074        255   147.9167
Protein             19    18.5333    19.5833
Calcium         43.963       18.8    75.4167
Iron            2.3815     3.2267      1.325



Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.5517

Attribute: Energy
Normal Distribution. Mean = 255 StdDev = 250
Attribute: Protein
Normal Distribution. Mean = 18.5333 StdDev = 250
Attribute: Calcium
Normal Distribution. Mean = 18.8 StdDev = 250
Attribute: Iron
Normal Distribution. Mean = 3.2267 StdDev = 250

Cluster: 1 Prior probability: 0.4483

Attribute: Energy
Normal Distribution. Mean = 147.9167 StdDev = 250
Attribute: Protein
Normal Distribution. Mean = 19.5833 StdDev = 250
Attribute: Calcium
```

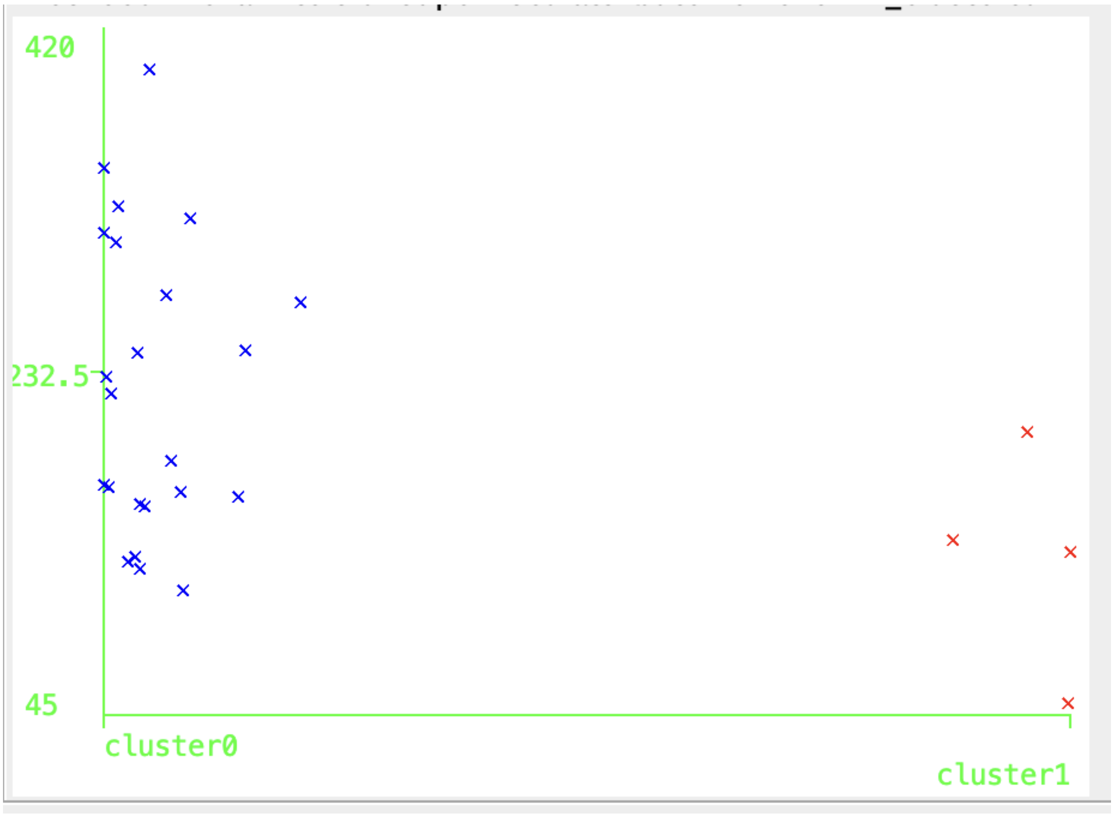


```
Normal Distribution. Mean = 75.4167 StdDev = 250
Attribute: Iron
Normal Distribution. Mean = 1.325 StdDev = 250


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

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

Clustered Instances

0       23 ( 85%)
1        4 ( 15%)


Log likelihood: -25.91152
```

Clusters according to Energy, minStdDev = 1e-6



Clusters according to Energy, minStdDev = 250



From running the MakeDensityBasedClusterer in two different standard deviation settings, we can see that by setting different minimum standard deviation limits, we try to define the minimum size of the Cluster. By comparing the 2 runs, we can see that the number of objects categorized as cluster #0 is 23 in the run with minStdDev = 250 while there are 15 objects in cluster #0 in the run with minStdDev = 1e-6.