Advanced Data Mining - Lab 1

Mim Kemal Tekin (mimte666)

1/28/2019

Contents

Lab 1: Clustering

Assignment 1: K-Means Algorithm

Apply "SimpleKMeans" to your data. In 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:

Task 1.1: Explanation of name attribute

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?)

"name" attribute should be ignored, because it is just a alias for tubles. Computer cannot get any information from "name" attribute to create clusters. But we can extract some knowledge from the names. In fact we can see that all the products are different type of meat and fish.

Task 1.2: K-Means k=2,3,4 with seed=10

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

In this task we tried to run k-means cluestering algorithm with $k = 2,\ 3,\ 5$ and $textseed = 10$. We can see the results of these 3 clustering results at following page. We can see the cluster centroids for each features and instance counts for each cluster.
When $k = 4$ cluster 3 has only one observation while others have many more observations. It suggests us that the optimal number of clusters should be between K=2 or 3. When $k = 2, 3$ we can see instances are distributed fairly between clusters.

Additionally, we can see also cluster plots after cluster info. It is clear to see, Energy vs Protein is clustered perfectly in both cases (K=2 and 3). When K=2 Energy vs Calsium and Energy vs Iron show us after 232.5 of Energy, Iron and Calcium are stable and it is reflected in the cluster composition. When K=3 instead, the third cluster is in fact a subset of the second cluster but without any clear boundary. Energy vs Fat highlights a small gap in the linear trend at energy = 232.5. When K=2 the difference is picked by the cluster division while with k=3 the 3rd cluster is overlaps with the second.

Cluster Info

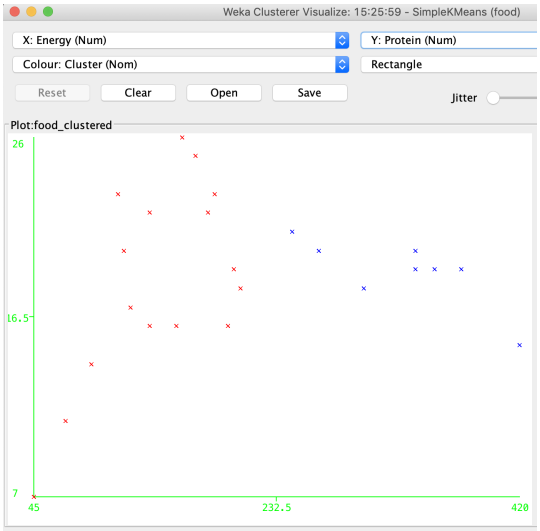
```
Scheme:weka.clusterers.SimpleKMeans -N 2 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     food
Instances:    27
Attributes:   6
              Energy
              Protein
              Fat
              Calcium
              Iron
Ignored:
              Name
Test mode:evaluate 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) : 0.01 seconds

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

Clustered Instances

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

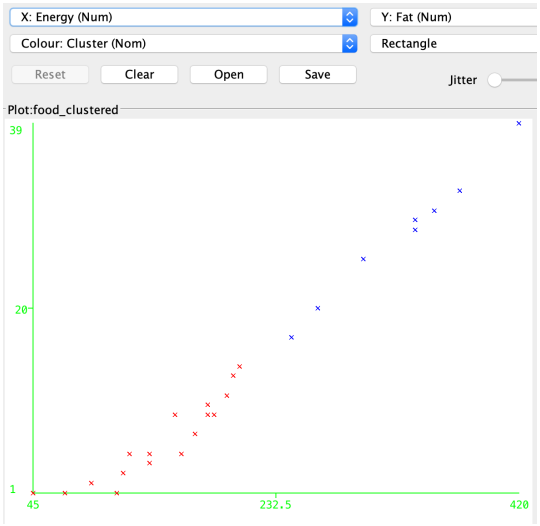
```
Scheme:weka.clusterers.SimpleKMeans -N 3 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     food
Instances:    27
Attributes:   6
              Energy
              Protein
              Fat
              Calcium
              Iron
Ignored:
              Name
Test mode:evaluate on training data

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


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 4.077107847192327
Missing values globally replaced with mean/mode

Cluster centroids:
                           Cluster#
Attribute    Full Data          0          1          2
                  (27)        (8)       (12)        (7)
=====================================================
Energy         207.4074    341.875     171.25   115.7143
Protein              19      18.75    22.1667    13.8571
Fat             13.4815     28.875       8.25     4.8571
Calcium          43.963       8.75    48.1667         77
Iron             2.3815     2.4375     2.3583     2.3571



Time taken to build model (full training data) : 0.01 seconds

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

Clustered Instances

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

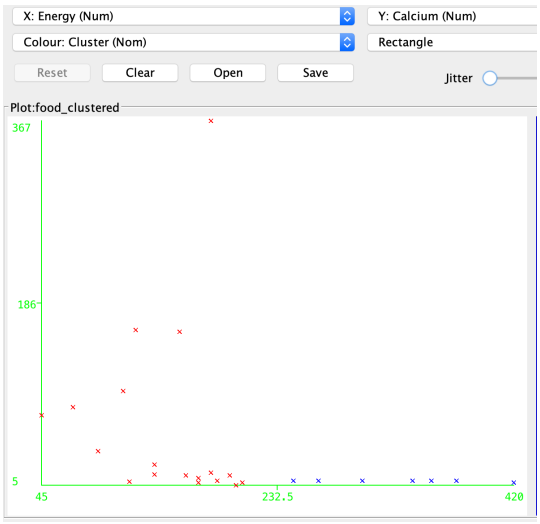
```
Scheme:weka.clusterers.SimpleKMeans -N 4 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     food
Instances:    27
Attributes:   6
              Energy
              Protein
              Fat
              Calcium
              Iron
Ignored:
              Name
Test mode:evaluate on training data

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


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 3.229030897655616
Missing values globally replaced with mean/mode

Cluster centroids:
                              Cluster#
Attribute    Full Data          0          1          2          3
                  (27)        (8)       (11)        (7)        (1)
============================================================================
Energy         207.4074    341.875   170.4545   115.7143        180
Protein              19      18.75    22.1818    13.8571         22
Fat             13.4815     28.875     8.1818     4.8571          9
Calcium          43.963       8.75    19.1818         77        367
Iron             2.3815     2.4375     2.3455     2.3571        2.5



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

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

Clustered Instances

0       8 ( 30%)
1      11 ( 41%)
2       7 ( 26%)
3       1 (  4%)
```

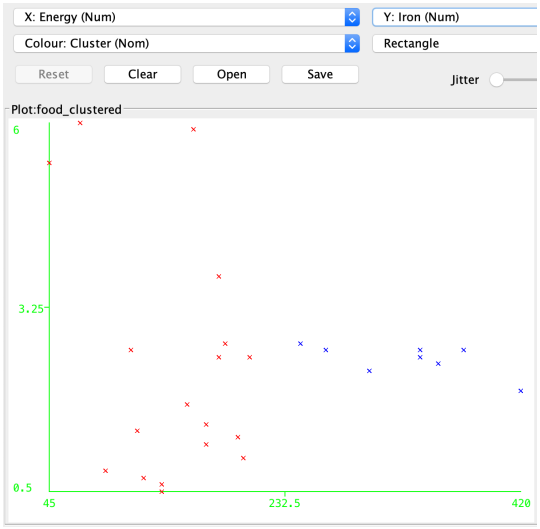
Cluster Plots

k=2

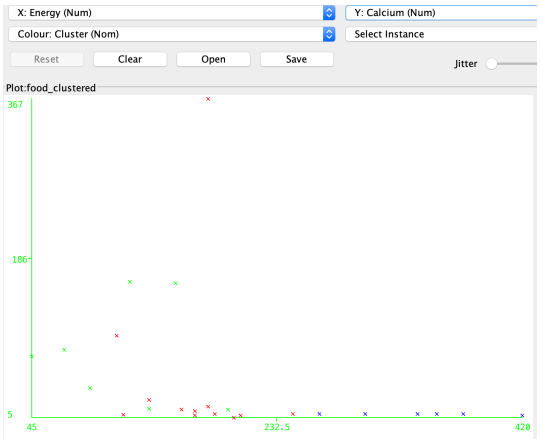


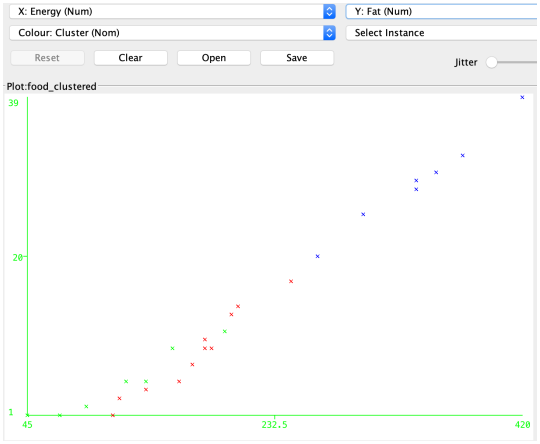


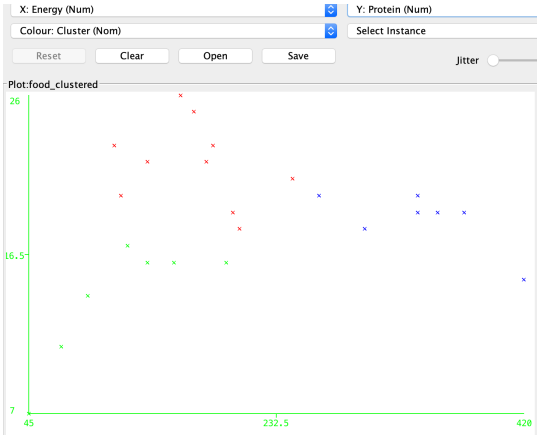


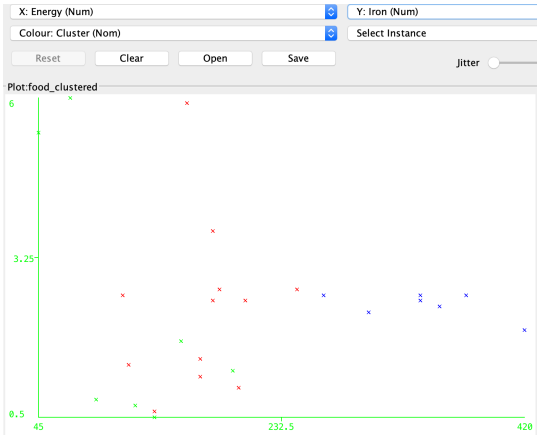


k=3


X: Energy (Num)
Y: Calcium (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:food_clustered
367
186
5
45
232.5
420


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: Energy (Num)
Y: Protein (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:food_clustered
26
16.5
7
45
232.5
420


X: Energy (Num)
Y: Iron (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:food_clustered
6
3.25
0.5
45
232.5
420

Task 1.3: K-Means k=2,3,4 with seed=666

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.

In this task we used seed = 666 and we got the result are following. This time when $k = 4$ we got 2 instances in cluster 2 as smallest cluster. It is better than seed = 10 but still it has way less observations than other clusters we got. We cannot see big changes when $k = 2$, but when $k = 3$ the smallest clusters have 6 units of difference while in the previus seed the difference is only one. Changing the seed does not influence k=2, since the swapped observation lies on the boundary of the clusters. We have only 2 outliers when K=3 and we can say that it is separated better than seed = 10.

The seed parameter influences the starting centroids, because it controls the random number generation. Having a different starting points lead to different clusters if there is no clear boundary between them.

Cluster Info

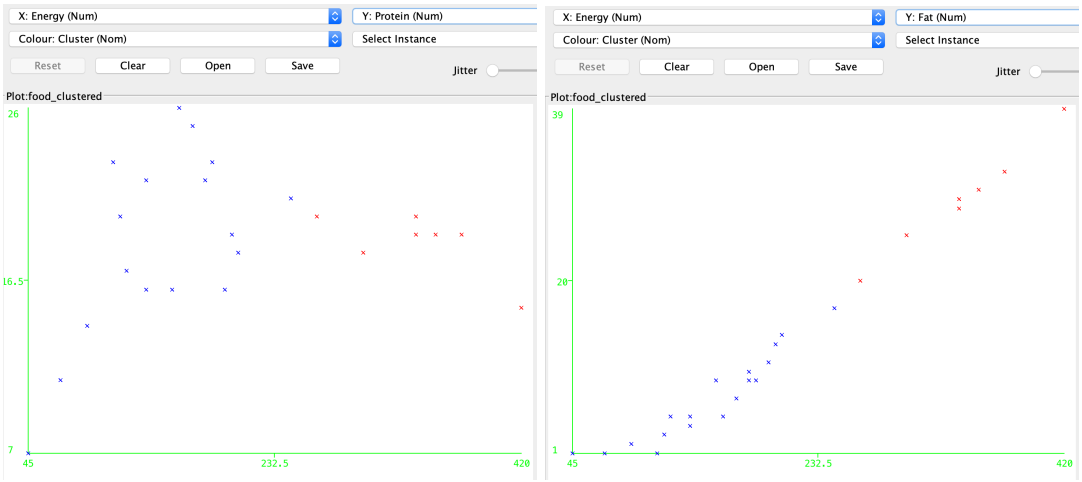
```
Scheme:weka.clusterers.SimpleKMeans -N 2 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -S
666
Relation:     food
Instances:    27
Attributes:   6
              Energy
              Protein
              Fat
              Calcium
              Iron
Ignored:
              Name
Test mode:evaluate on training data

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


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 5.0829748461313
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
                  (27)       (19)        (8)
=============================================
Energy        207.4074   150.7895    341.875
Protein             19    19.1053      18.75
Fat            13.4815          7     28.875
Calcium         43.963    58.7895       8.75
Iron            2.3815     2.3579     2.4375



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

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

Clustered Instances

0      19 ( 70%)
1       8 ( 30%)
```

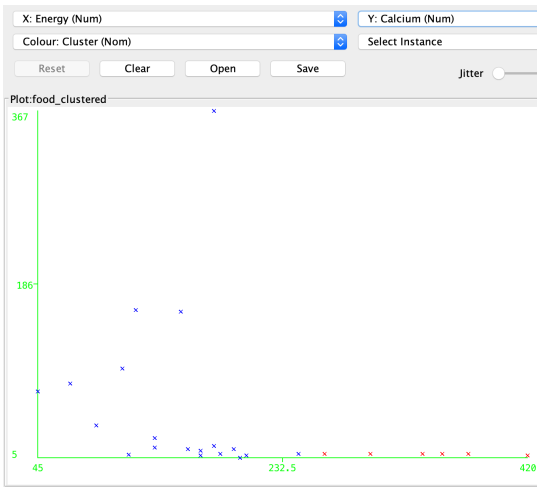
```
Scheme:weka.clusterers.SimpleKMeans -N 3 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -S 666
Relation:     food
Instances:    27
Attributes:   6
              Energy
              Protein
              Fat
              Calcium
              Iron
Ignored:
              Name
Test mode:evaluate on training data

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


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 3.424944597251442
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute    Full Data          0          1          2
                  (27)       (17)        (8)        (2)
=====================================================
Energy        207.4074   161.7647    341.875       57.5
Protein             19    20.2941      18.75          9
Fat            13.4815     7.7059     28.875          1
Calcium         43.963    56.5294       8.75         78
Iron            2.3815     1.9647     2.4375        5.7



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

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

Clustered Instances

0      17 ( 63%)
1       8 ( 30%)
2       2 (  7%)
```

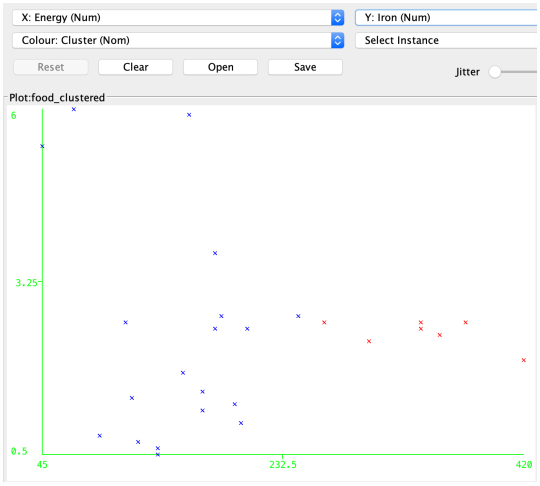
```
Scheme:weka.clusterers.SimpleKMeans -N 4 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -S 666
Relation:     food
Instances:    27
Attributes:   6
              Energy
              Protein
              Fat
              Calcium
              Iron
Ignored:
              Name
Test mode:evaluate on training data

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


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 3.0368466602673103
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3
                  (27)       (10)        (7)        (2)        (8)
=============================================================================
Energy          207.4074      185.5   352.8571       57.5        145
Protein               19       18.5    18.5714          9       22.5
Fat              13.4815         11    30.1429          1      5.125
Calcium           43.963         28     8.7143         78      86.25
Iron              2.3815       1.96     2.4143        5.7       2.05



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

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

Clustered Instances

0      10 ( 37%)
1       7 ( 26%)
2       2 (  7%)
3       8 ( 30%)
```

Cluster Plots

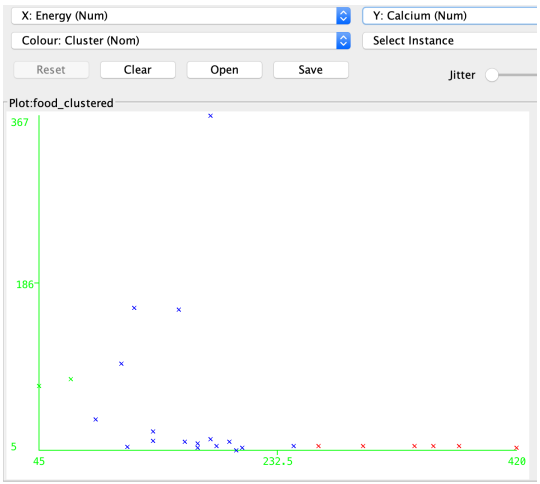
k=2

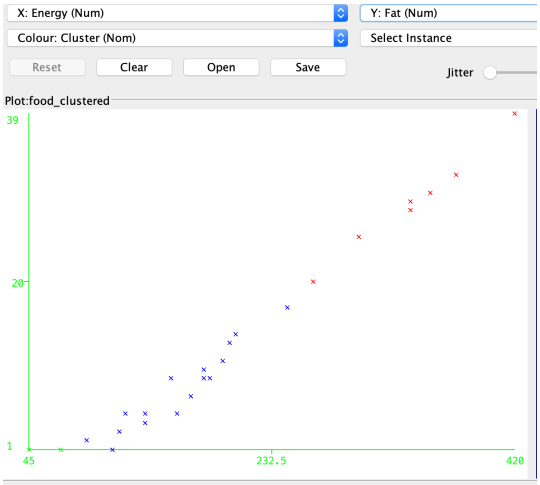


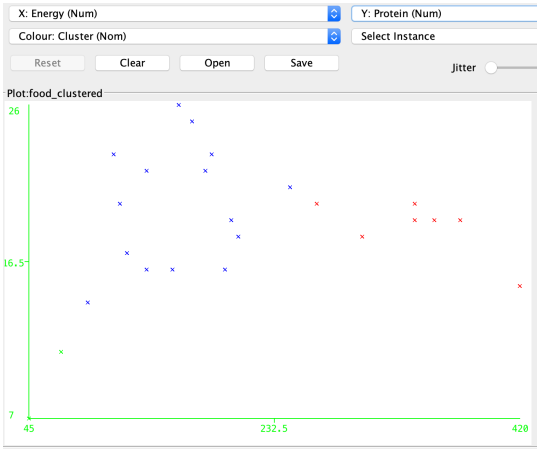

X: Energy (Num)
Y: Calcium (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:food_clustered
367
186
5
45
232.5
420

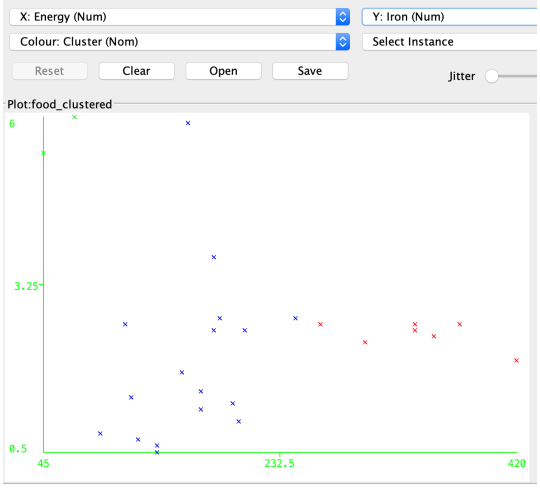

X: Energy (Num)
Y: Iron (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:food_clustered
6
3.25
0.5
45
232.5
420

k=3


X: Energy (Num)
Y: Calcium (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:food_clustered
367
186
5
45
232.5
420


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: Energy (Num)
Y: Protein (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:food_clustered
26
16.5
7
45
232.5
420


X: Energy (Num)
Y: Iron (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:food_clustered
6
3.25
0.5
45
232.5
420

Task 1.4: Cluester Analysis

Do you think the clusters are "good" clusters?

Yes, they are "good" clusters as we discussed in the previous tasks before. When K=2 we have 2 clusters which are separated well. When K=3 we have 2 cluster which are mixed each other, but still it is separeted with the 3rd cluster.

Task 1.5: Name of Clusters

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

As we can see the units of the first cluster are all belonging to red meat and the other cluster instead is made of white meat. We can see the relations in the cluster plots also. After a short research we can say the values of features also prove the result.

```
1 0 | Braised beef
1 0 | Hamburger
1 0 | Roast beef
1 0 | Beefsteak
0 1 | Canned beef
0 1 | Broiled chicken
0 1 | Canned chicken
0 1 | Beef heart
1 0 | Roast lamb leg
1 0 | Roast lamb shoulder
1 0 | Smoked ham
1 0 | Pork roast
1 0 | Pork simmered
0 1 | Beef tongue
0 1 | Veal cutlet
0 1 | Baked bluefish
0 1 | Raw clams
0 1 | Canned clams
0 1 | Canned crabmeat
0 1 | Fried haddock
0 1 | Broiled mackerel
0 1 | Canned mackerel
0 1 | Fried perch
0 1 | Canned salmon
0 1 | Canned sardines
0 1 | Canned tuna
0 1 | Canned shrimp
```

Assignment 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)

```
1 0 | Braised beef
1 0 | Hamburger
1 0 | Roast beef
1 0 | Beefsteak
0 1 | Canned beef
0 1 | Broiled chicken
0 1 | Canned chicken
0 1 | Beef heart
1 0 | Roast lamb leg
1 0 | Roast lamb shoulder
1 0 | Smoked ham
1 0 | Pork roast
1 0 | Pork simmered
1 0 | Beef tongue
0 1 | Veal cutlet
0 1 | Baked bluefish
0 1 | Raw clams
0 1 | Canned clams
0 1 | Canned crabmeat
0 1 | Fried haddock
0 1 | Broiled mackerel
0 1 | Canned mackerel
0 1 | Fried perch
0 1 | Canned salmon
0 1 | Canned sardines
0 1 | Canned tuna
0 1 | Canned shrimp
```

```
0 1 | Braised beef
0 1 | Hamburger
1 0 | Roast beef
1 0 | Beefsteak
0 1 | Canned beef
0 1 | Broiled chicken
0 1 | Canned chicken
0 1 | Beef heart
0 1 | Roast lamb leg
0 1 | Roast lamb shoulder
0 1 | Smoked ham
0 1 | Pork roast
0 1 | Pork simmered
0 1 | Beef tongue
0 1 | Veal cutlet
0 1 | Baked bluefish
0 1 | Raw clams
0 1 | Canned clams
0 1 | Canned crabmeat
0 1 | Fried haddock
0 1 | Broiled mackerel
0 1 | Canned mackerel
0 1 | Fried perch
0 1 | Canned salmon
0 1 | Canned sardines
0 1 | Canned tuna
0 1 | Canned shrimp
```

First image above is $sd = 10^{-6}$ and the second one is $sd = 2 * 10^2$. In the first clusters we can see there is only one difference: "Beef tongue" passes from White Meat cluster to Red Meat cluster which is an improvement. When we increase the sd to $2 * 10^2$, many red meats passes to the white meat cluster. To high standard deviation results in flat densities which leads to the unification of the clusters. In fact a high variance will result into a really flat density, which means that the difference in density between the observations on the tails and the observations close to the mean will be really small, making the identification of clusters more and more difficult.