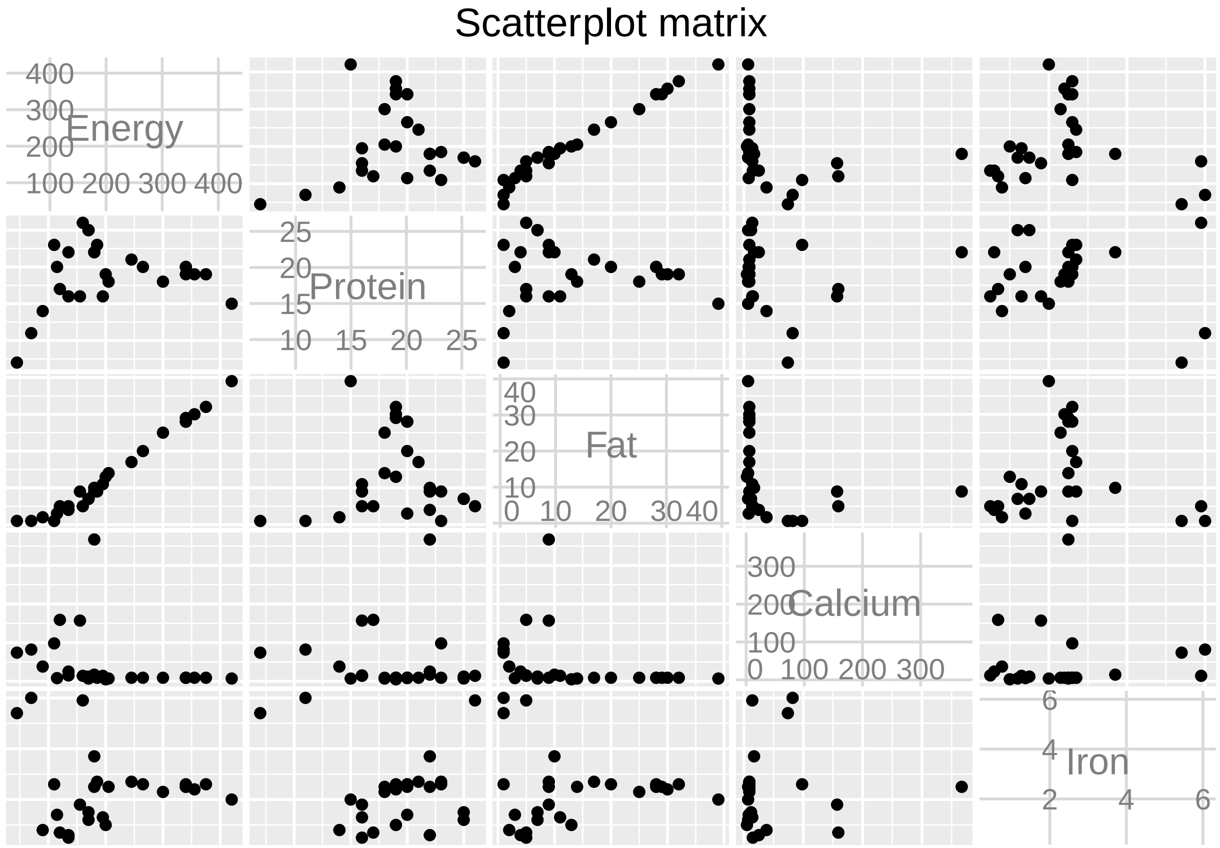
Data Mining - Lab 1 - Cluster Analysis

Gustav Sternelöv

3 februari 2016

Simple K-means

I start with making a scatterplot which includes all variables except *names*. The latter is excluded because it just gives the names of the respective products. An alternative could have been to create a categorical variable since the names of the products indicates quite well what kind of food it contains. However, K-means is not a good algorithm for categorical data so the variable would still not have been so interesting to include in this particular analysis.



The chosen attributes are *Energy*, *Protein*, *Fat* and *Iron*. This is motivated by the patterns visualised in the scatterplot. My conclusion is that all of the variables but *calcium* seem to be interesting to include. For *calcium* most of the values are very close to each other apart from some outliers. It is therefore interpreted as not being the most interesting variable to include. The other variables have values that are more spread out in different groups which might be possible to investigate closer with a cluster analysis.

Seed 10

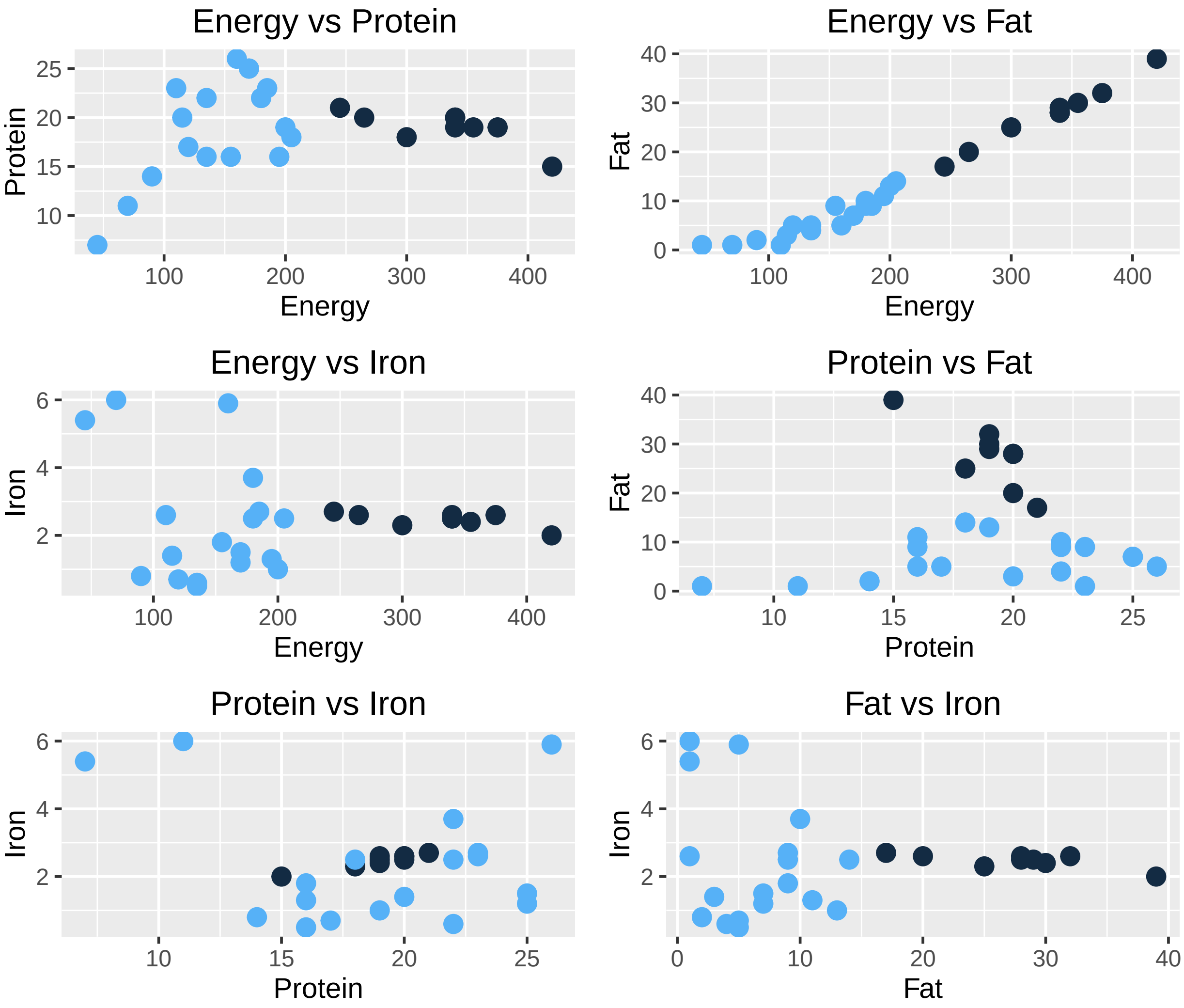
The initial clusters centroids are held fixed with the seed value *10* and two different cluster analysis are performed. In the first case the data points are divided into two clusters and in the second case into four different clusters.
Use K=3 or K=4 in the second case?

Two clusters

The number of iterations is two and the within cluster *SSE* is 3.99. The first cluster contains *1/3* of the data points and the second cluster the remainder of the data points. How the initial starting points has changed can be seen in the output below.

```
##
## kMeans
## ======
##
## Number of iterations: 2
## Within cluster sum of squared errors: 3.9886919330126585
##
## Initial starting points (random):
##
## Cluster 0: 340,20,28,2.6
## Cluster 1: 170,25,7,1.5
##
## Missing values globally replaced with mean/mode
##
## Final cluster centroids:
##                          Cluster#
## Attribute    Full Data          0          1
##                 (27.0)      (9.0)     (18.0)
## =============================================
## 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
```

Another way to present the obtained clusters are through visualisation. How the data points for each variable are clustered into the respective clusters is shown by the graphs below.
Exclude iron or not?

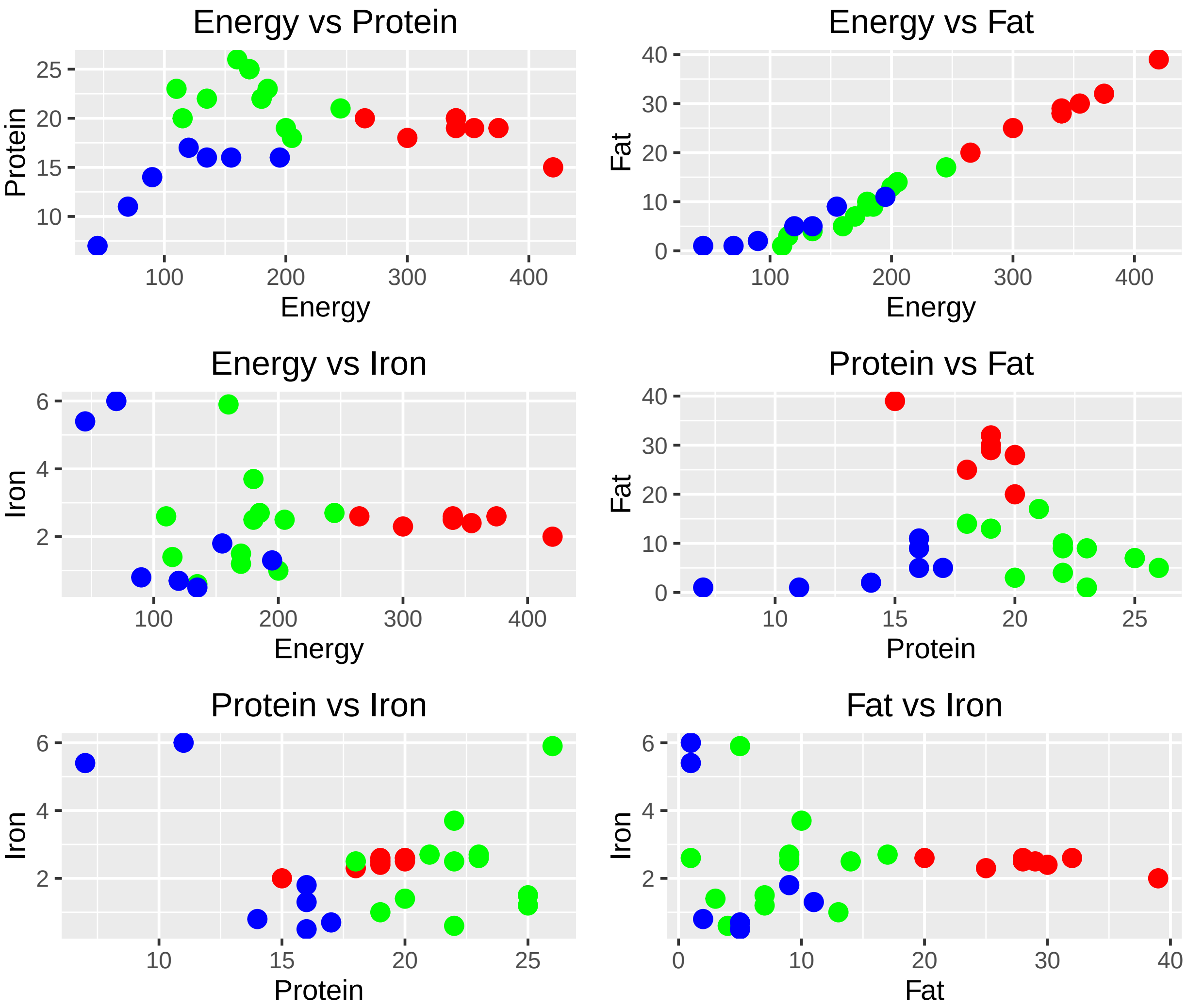


Four clusters

For the clustering with K=... the number of iterations is ... and the SSE is ... . Information about gow many of the points that are clustered into each cluster and how the initial staring points has changed is given in the output below.

```
##
## kMeans
## ======
##
## Number of iterations: 3
## Within cluster sum of squared errors: 3.0045584235890583
##
## Initial starting points (random):
##
## Cluster 0: 340,20,28,2.6
## Cluster 1: 170,25,7,1.5
## Cluster 2: 90,14,2,0.8
##
```

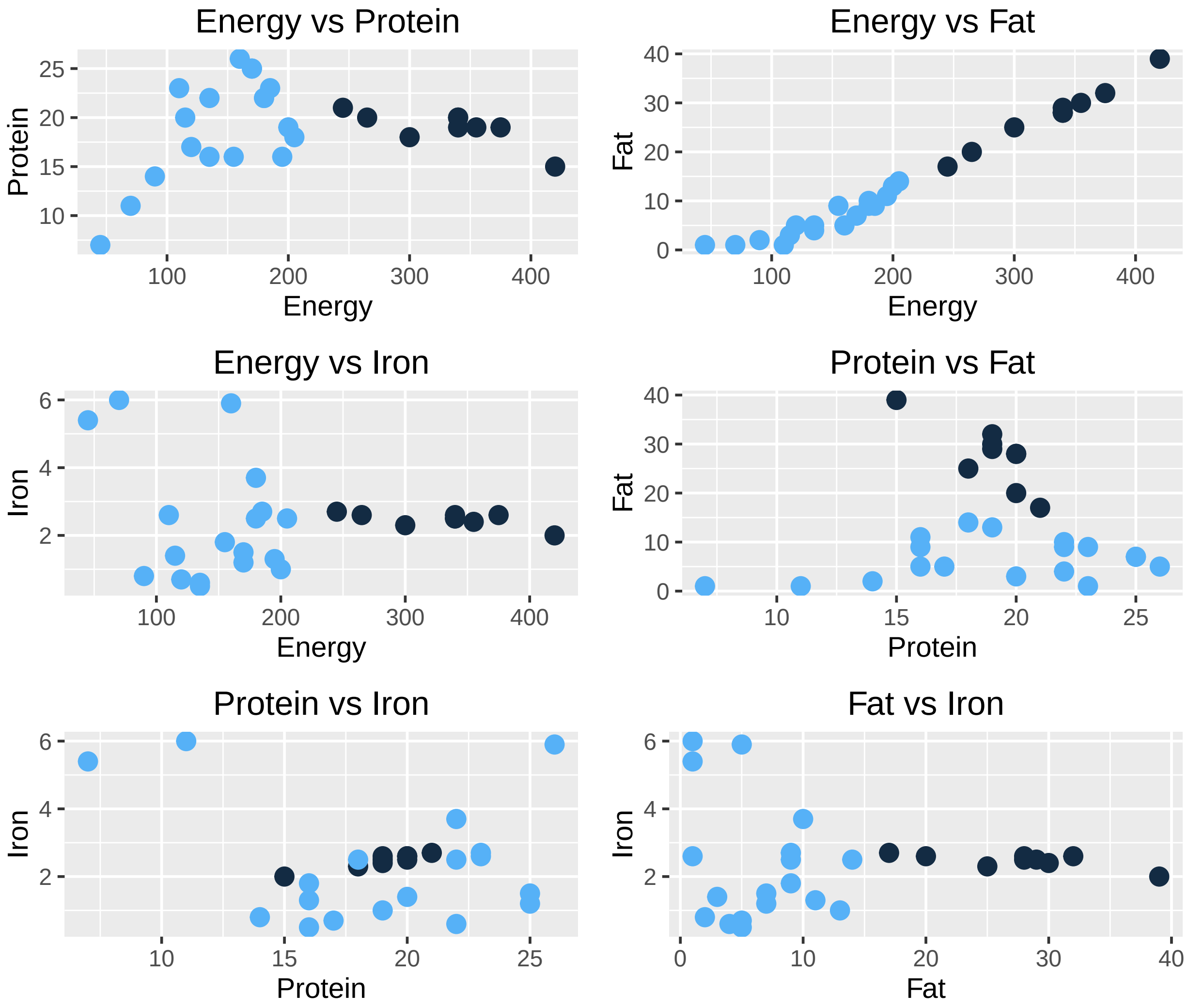
```
## Missing values globally replaced with mean/mode
##
## Final cluster centroids:
##                           Cluster#
## Attribute     Full Data          0          1          2
##                  (27.0)      (8.0)     (12.0)      (7.0)
## =========================================================
## 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
## Iron             2.3815     2.4375     2.3583     2.3571
```



Seed 28

Two clusters

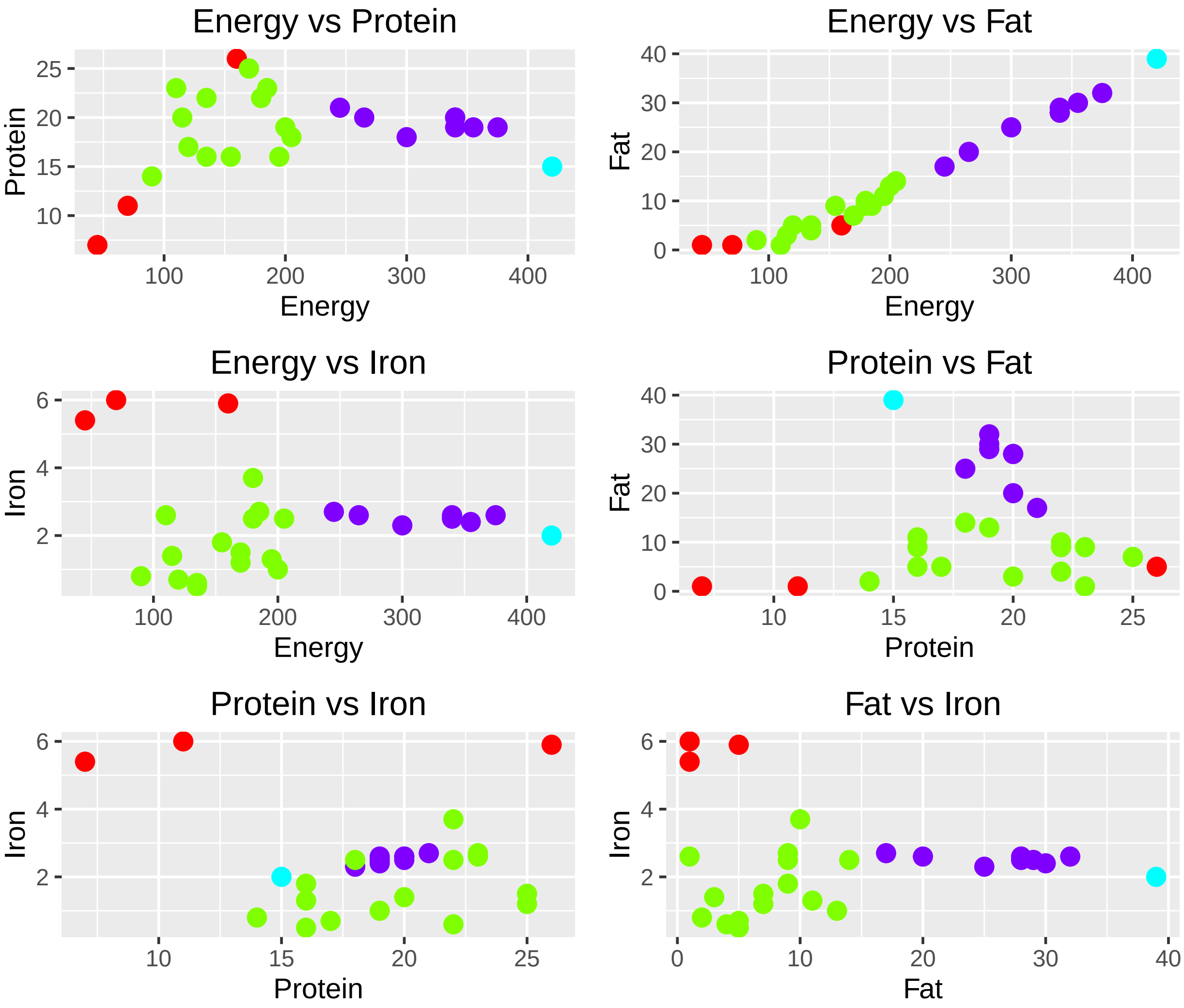
```
##
## kMeans
## ======
##
## Number of iterations: 6
## Within cluster sum of squared errors: 3.9886919330126585
##
## Initial starting points (random):
##
## Cluster 0: 155,16,9,1.8
## Cluster 1: 90,14,2,0.8
##
## Missing values globally replaced with mean/mode
##
## Final cluster centroids:
##                          Cluster#
## Attribute    Full Data          0          1
##                 (27.0)      (9.0)     (18.0)
## ============================================
## 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
```



Four clusters

```
##
## kMeans
## ======
##
## Number of iterations: 8
## Within cluster sum of squared errors: 2.0634372954170264
##
## Initial starting points (random):
##
## Cluster 0: 155,16,9,1.8
## Cluster 1: 90,14,2,0.8
## Cluster 2: 420,15,39,2
## Cluster 3: 340,20,28,2.6
##
## Missing values globally replaced with mean/mode
##
## Final cluster centroids:
```

```
##                          Cluster#
## Attribute    Full Data          0          1          2          3
##                (27.0)      (3.0)     (15.0)      (1.0)      (8.0)
## =================================================================
## Energy        207.4074    91.6667   156.3333        420        320
## Protein             19    14.6667    19.8667         15       19.5
## Fat            13.4815     2.3333     7.2667         39     26.125
## Iron            2.3815     5.7667     1.6533          2      2.525
```



MakeDensityBasedClusters