Advanced Data Mining Assignment 1: Cluster Analysis

Nazli Bilgic (nazbi056) & Simge Cinar (simci637)

February 8^{th} 2024

1 SimpleKmeans

The food data has 27 rows with 6 attributes which are Name, Energy, Protein, Fat, Calcium and Iron. Name attribute is string and the rest of the attributes are numeric. First an exploratory analysis is conducted using R and Weka. From the Figure 3, it was observed that Calcium has outlier values.

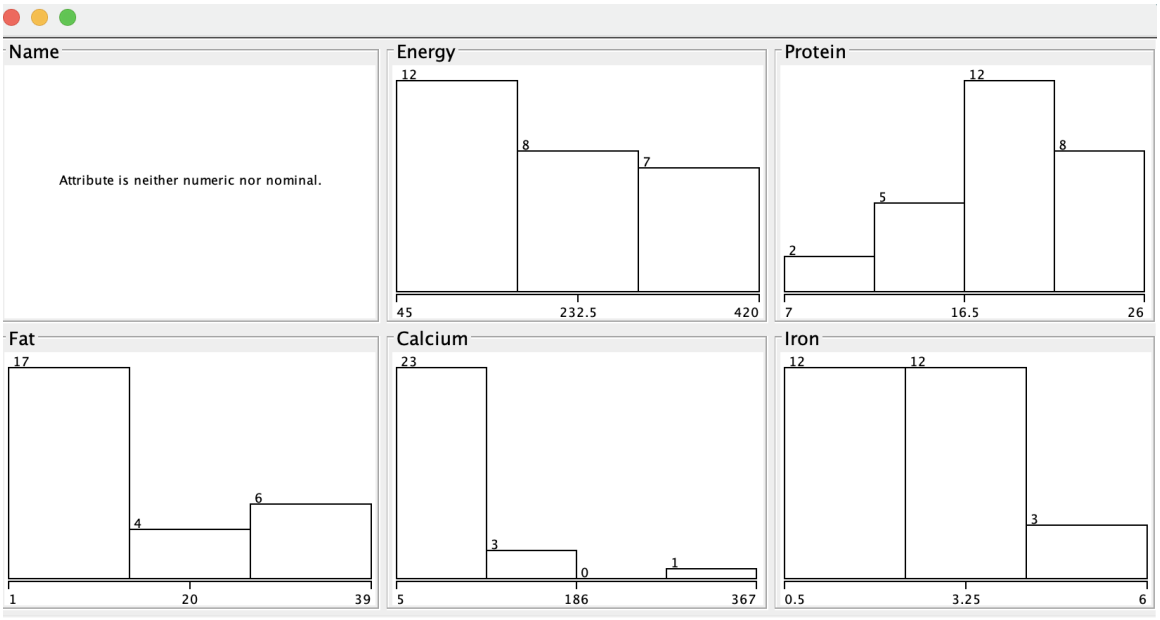

Figure 1: Histogram of the data

```
    Name              Energy          Protein          Fat            Calcium           Iron
 Length:27         Min.   : 45.0   Min.   : 7.0   Min.   : 1.00   Min.   :  5.00   Min.   :0.500
 Class :character  1st Qu.:135.0   1st Qu.:16.5   1st Qu.: 5.00   1st Qu.:  9.00   1st Qu.:1.350
 Mode  :character  Median :180.0   Median :19.0   Median : 9.00   Median :  9.00   Median :2.500
                   Mean   :207.4   Mean   :19.0   Mean   :13.48   Mean   : 43.96   Mean   :2.381
                   3rd Qu.:282.5   3rd Qu.:22.0   3rd Qu.:22.50   3rd Qu.: 31.50   3rd Qu.:2.600
                   Max.   :420.0   Max.   :26.0   Max.   :39.00   Max.   :367.00   Max.   :6.000
```

Figure 2: Summary of the data

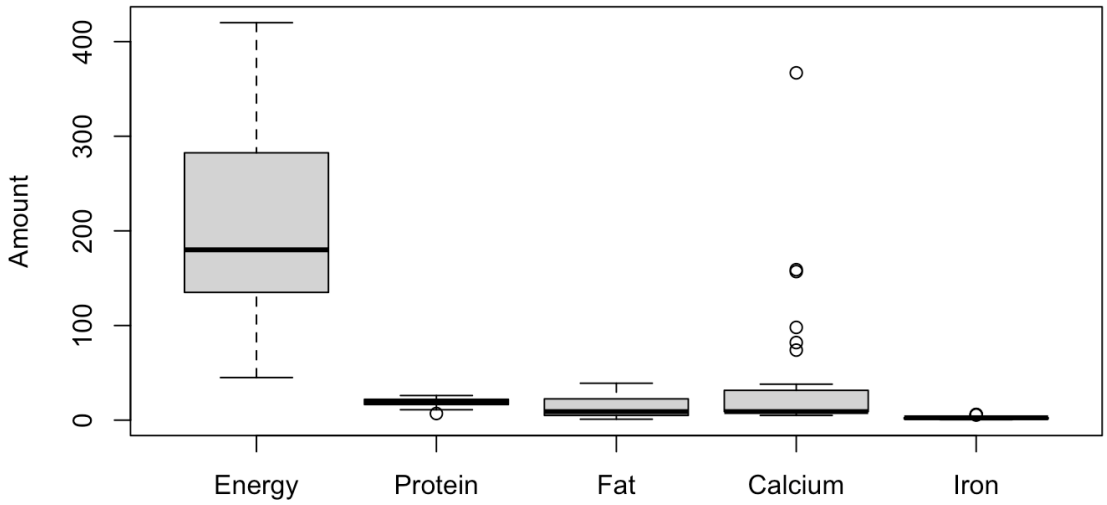

Figure 3: Boxplot

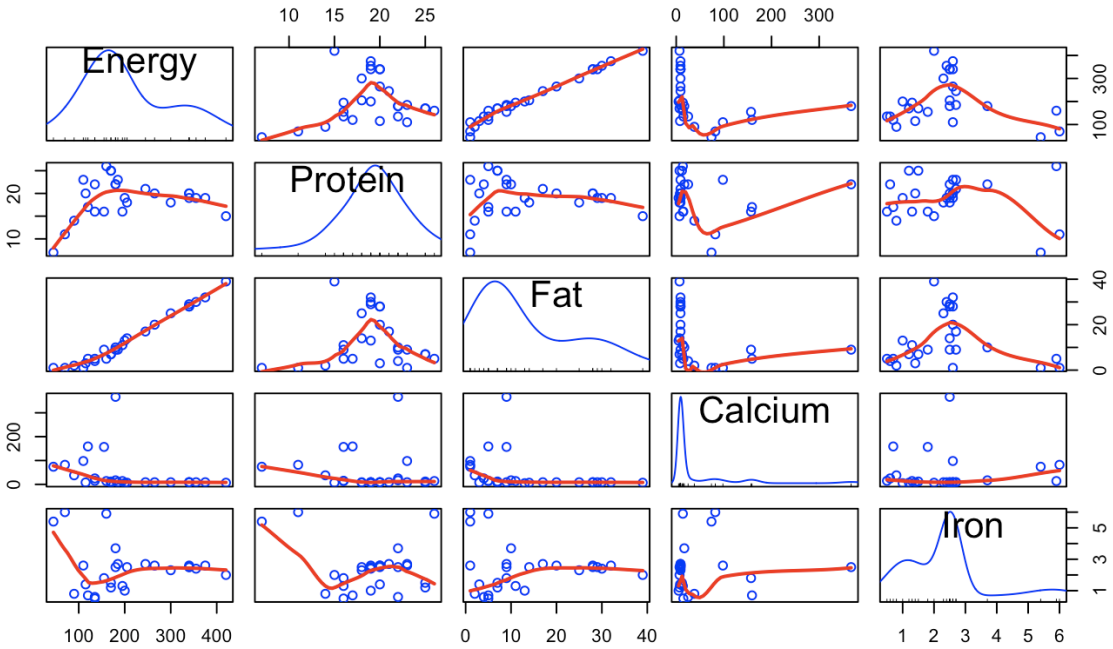

Figure 4: Scatterplot

1.1 Question 1

From the Figure 4, it can be observed that there is a positive correlation between "Fat" and "Energy", the correlation coefficient is 0.9871. Considering the context of the data, they contain similar information hence "Energy" is excluded to avoid redundancy. Moreover, "Name" attribute is excluded since it is a string. String attributes do not have meaningful distance measure. Overall, "Protein", "Fat", "Calcium" and "Iron" attributes were used the for k-means algorithm.

1.2 Question 2

Sum of squared values are calculated with seed 10 for each cluster from 2 to 11. k = 5 and k = 6 is selected using elbow method. The SSE values can be seen in the Figure 5 for each cluster. The rate of decrease sharply changes when k = 5. Also k = 6 is chosen since they have similar errors with k = 5. The results of the k-means algorithm can be seen in the Figure 6 and 7 for k = 5 and k = 6 respectively.

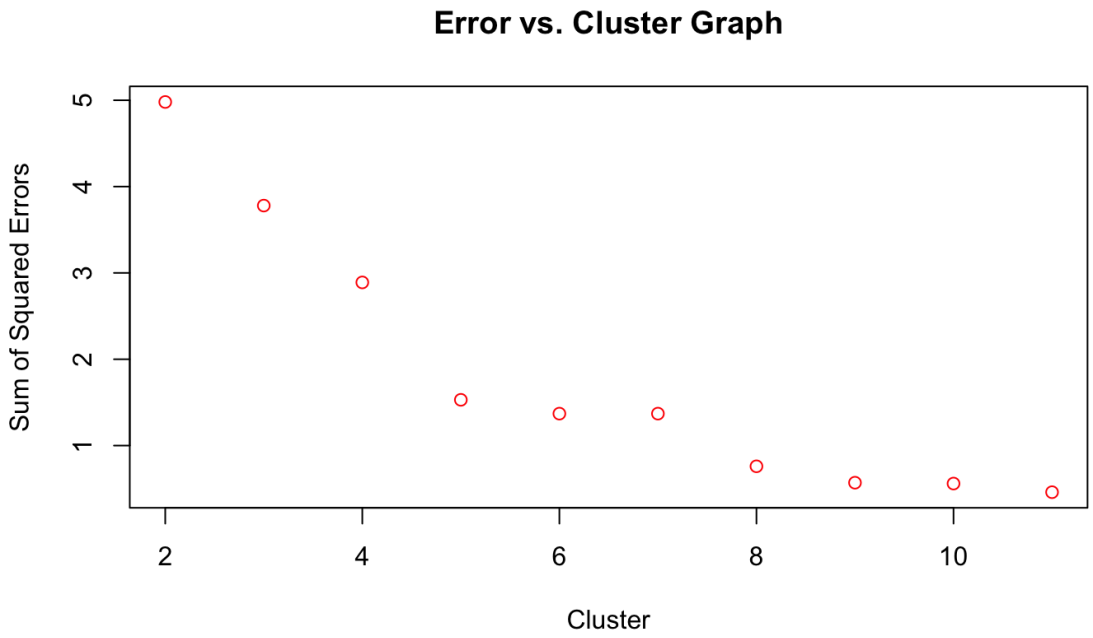

Figure 5: Error vs. Cluster Analysis

```
Number of iterations: 9
Within cluster sum of squared errors: 1.5306985143106107
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4
                  (27)        (8)        (7)        (2)        (1)        (9)
=============================================================================
Protein             19      18.75    23.5714          9         22    17.5556
Fat            13.4815     28.875          8          1          9     7.3333
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) : 0.01 seconds
```

Figure 6: k = 5, seed = 10

When k = 5 and seed = 10,

- Protein is highest in cluster 1 and lowest in cluster 2. From this it can be observed that foods which are in cluster 1 are richer on Protein than the other clusters.
- Fat is highest in cluster 0 and lowest in cluster 2.
- Calcium is highest in cluster 3 and lowest in cluster 0. We can comment that cluster 3 has contains foods that are good sources of calcium.
- Iron is highest in cluster 2 and lowest in cluster 0. Cluster 2 has the foods that are good sources for iron.
- Cluster 0, 1 and 4 has balanced number instances whereas cluster 2 and 3 has relatively low instances.
- Sum of squared error: 1.5307

```
Number of iterations: 4
Within cluster sum of squared errors: 1.374684880533414
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4          5
                  (27)        (7)        (8)        (2)        (1)        (4)        (5)
=============================================================================================
Protein             19    18.5714      23.25          9         22       19.5       15.8
Fat            13.4815    30.1429       5.75          1          9         16        6.4
Calcium         43.963     8.7143      23.75         78        367        7.5       76.6
Iron            2.3815     2.4143       2.45        5.7        2.5        2.2       1.02



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

Figure 7: k = 6, seed = 10

When k = 6 and seed = 10,

- Protein is highest in cluster 1 and cluster 3 has close value to cluster 1, lowest in cluster 2. From this we can comment that foods which are in cluster 1 and 3 are richer on Protein than the other clusters.
- Fat is highest in cluster 0 and lowest in cluster 2.
- Calcium is highest in cluster 3 and lowest in cluster 4. We can comment that cluster 3 has contains foods that are good sources of calcium.
- Iron is highest in cluster 2 and lowest in cluster 5. Cluster 2 has the foods that are good sources for iron.
- Cluster 2 and 3 has relatively low instances.
- Sum of squared errors: 1.3746

Sum of squared error is lower in k = 5 compared to k = 6 with the same seed value as expected.

1.3 Question 3

SSE and number of iterations for different seed and k values can be seen in the table below. Seed value controls the initial location of the cluster centroids which can influence the final clustering outcome. The number of iterations and SSE change with different seed values as expected because k-means cannot find global optimum hence different initial values converges to different local optimum points.

k	seed	SSE	Iteration
5	10	1.5307	9
5	20	1.9571	3
5	30	1.8439	4
6	10	1.3747	4
6	20	1.3712	3
6	30	1.6078	6

1.4 Question 4

Food names for each cluster can be seen below for k = 5 and k = 6.

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Braised beef	Hamburger	Raw clams	Canned sardines	Broiled chicken
Roast beef	Canned beef	Canned clams		Beef tongue
Beefsteak	Canned chicken			Baked bluefish
Roast lamb leg	Beef heart			Canned crabmeat
Roast lamb shoulder	Veal cutlet			Fried haddock
Smoked ham	Canned tuna			Broiled mackerel
Pork roast	Canned shrimp			Canned mackerel
Pork simmered				Fried perch
				Canned salmon

Table 1: Data with 5 clusters, seed = 10

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Raw clams	Canned crabmeat	Hamburger	Braised beef	Canned sardines	Canned chicken
Canned clams	Fried haddock	Canned beef	Roast beef	Canned shrimp	Broiled chicken
	Canned mackerel	Beef heart	Beefsteak		Broiled mackerel
	Fried perch	Roast lamb leg	Roast lamb shoulder		Baked bluefish
	Canned salmon	Beef tongue	Smoked ham		Canned tuna
		Veal cutlet	Pork roast		
			Pork simmered		

Table 2: Data with 6 clusters, seed = 20

The clusters are good clusters, especially for k = 5. Properties for clusters when k = 5 and seed = 10 is as follows:

- Cluster 0 contains the food that has high fat.
- The foods in cluster 1 are rich in terms of protein and poor in terms of fat.
- Cluster 2 contains sea food which has high calcium and high iron but low protein and fat.
- Cluster 3 has only 1 food which is canned sardines and it has the highest calcium rate among others. That calcium value can be considered as outlier.
- Cluster 4 has the food that are rich in terms of protein and low fat values but those food has higher calcium compared to cluster 2. They are generally fish.

Moreover it can be concluded that canned and broiled foods have low fat values. Also sea foods have high calcium.

When k = 6, the clusters are not that much good. For example roast lamb leg in cluster 2 and roast lamb shoulder in cluster 3 has similar attributes but they are not in the same cluster.

1.5 Question 5

Cluster 5 with seed value 10 is chosen. The labels for each cluster is as follows:
Cluster 0: Red meat with high fat
Cluster 1: High protein, low fat foods
Cluster 2: Clams
Cluster 3: Sardines (might be outlier)
Cluster 4: High calcium sea food
Cluster 5: White meat

2 MakeDensityBasedClusters

MakeDensityBasedClusterer fits a symmetric normal distribution to each cluster and minStdDev parameter controls the minimum standard deviation of the normal distribution in each cluster. In the first question, k = 5 with seed = 10 is selected as the best clustering. The number of instances with each minimum standard deviation can be seen in the table below. It can observed that increasing minStdDev creates larger clusters but excessively large standard deviations can result in clusters that are too spread out, like when minStdDev is 1000.

minStdDev	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1.0E-6	10	7	2	1	7
0.001	10	7	2	1	7
1	8	8	2	1	8
10	7	13	0	1	6
100	1	0	0	1	25
1000	0	0	0	0	27