**LAPORAN UJIAN AKHIR SEMESTER**

**STATISTIKA DESKRIPTIF**

**CLUSTERING**



**NAMA : MUKHAMAD IKHSANUDIN**

**NIM : 082011633086**

**DOSEN PENGAMPU : Drs. ETO WURYANTO, DEA.**

**196609281991021001**

**PROGRAM STUDI S1 SISTEM INFORMASI**

**FAKULTAS SAINS DAN TEKNOLOGI**

**UNIVERSITAS AIRLANGGA**

**2021**

```{R}

# Clustering

library(flexclust)

library(cluster)

library(factoextra)

library(mclust)

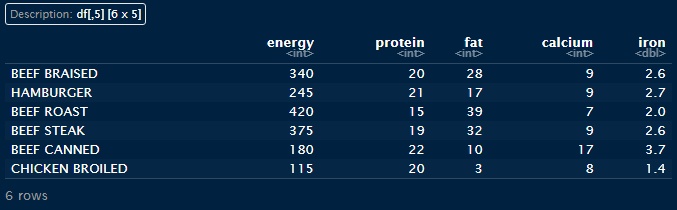
library(Gmedian)

# 1. Dataset Preparation

data("nutrient")

DataClust <- nutrient

Head(DataClust)



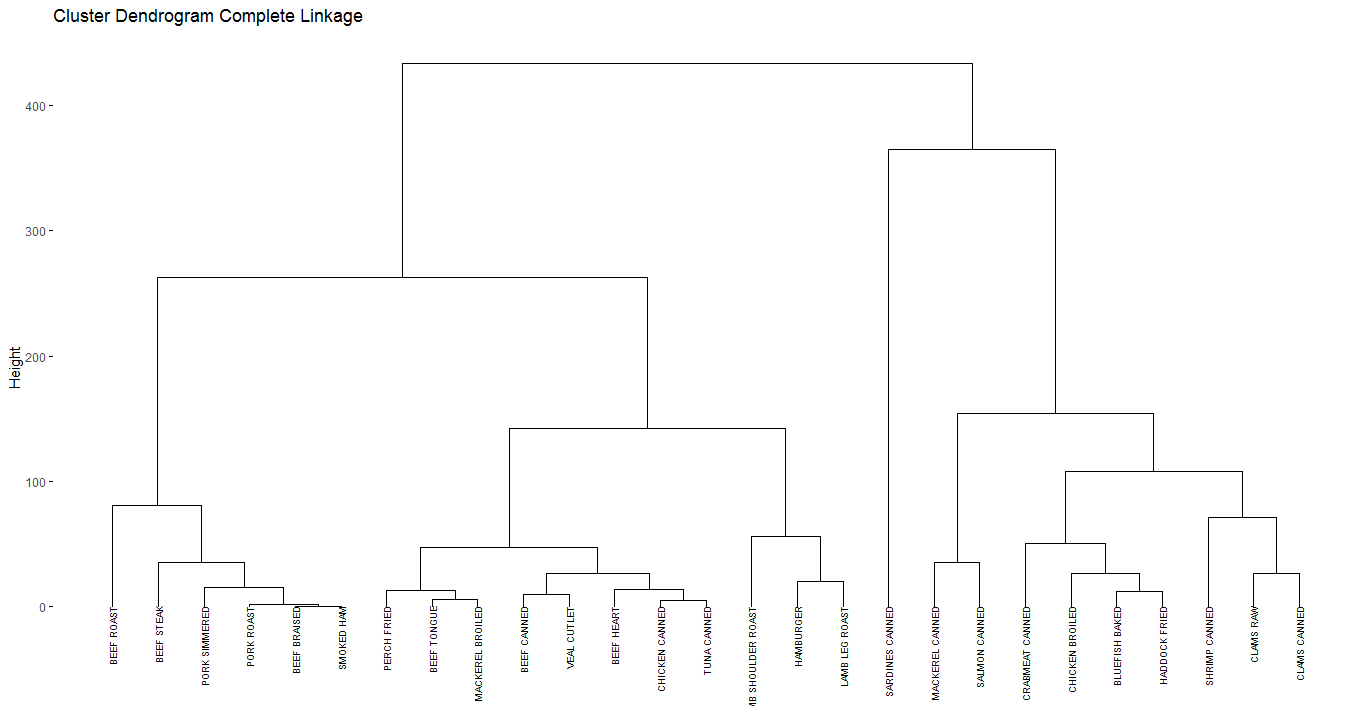
# Metode Agglomerative

# Complete Linkage

Clust\_Com <- hclust(dist(DataClust), method = "complete")

Clust\_Com

fviz\_dend(Clust\_Com, cex = 0.4, main = "Cluster Dendrogram Complete Linkage")

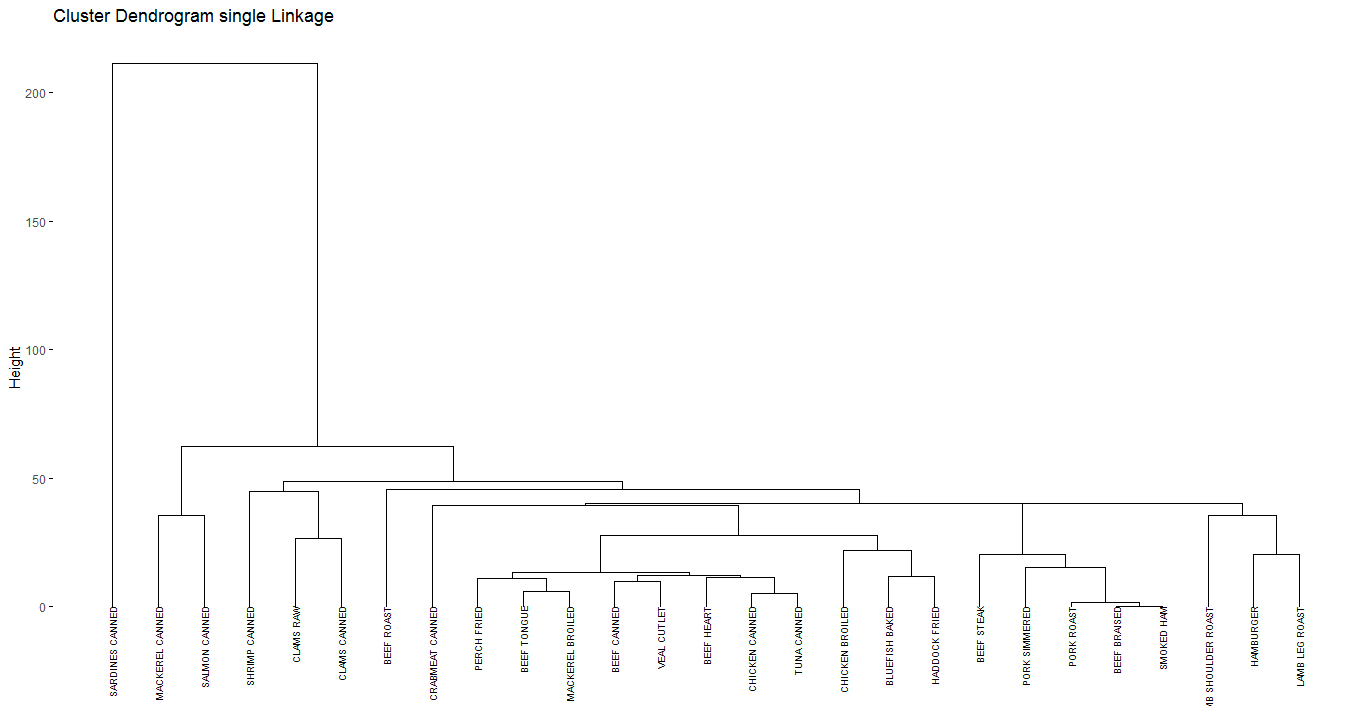


# Single Linkage

Clust\_Sin <- hclust(dist(DataClust), method = "single")

Clust\_Sin

fviz\_dend(Clust\_Sin, cex = 0.4, main = "Cluster Dendrogram single Linkage")

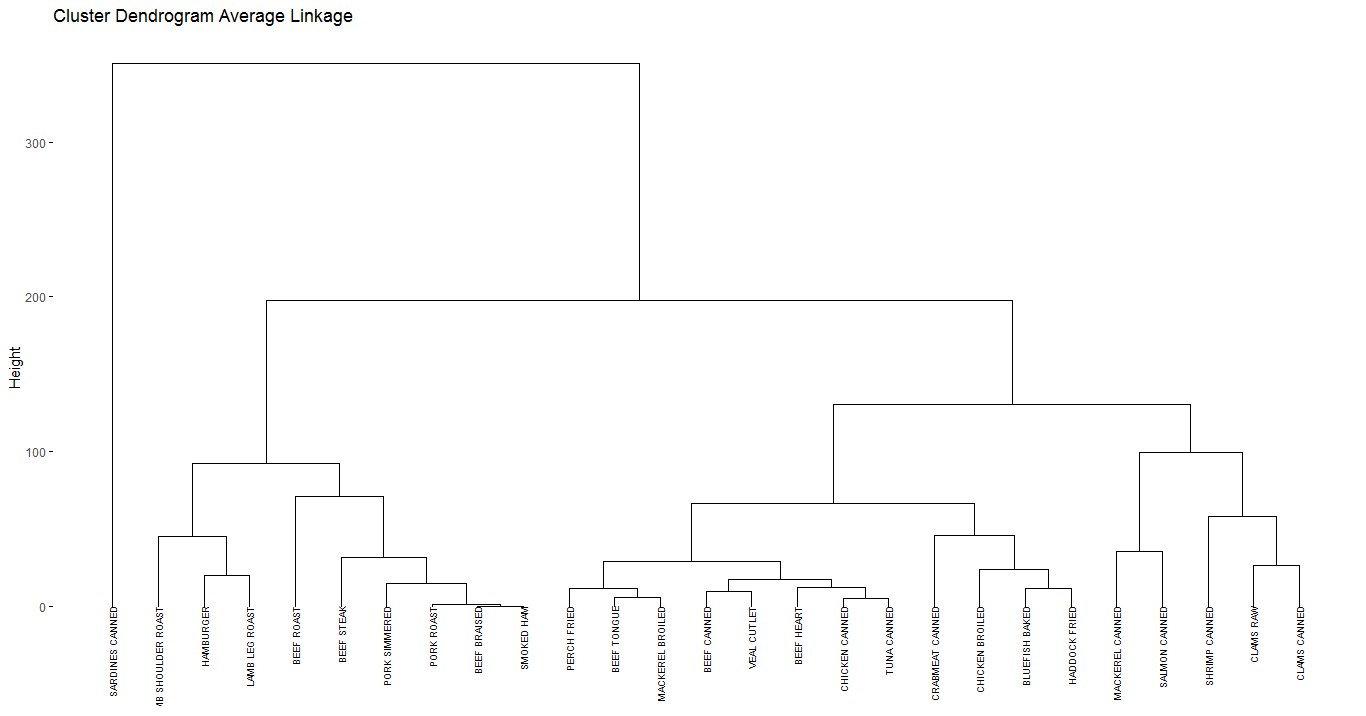


# Average Linkage

Clust\_Ave <- hclust(dist(DataClust), method = "average")

Clust\_Ave

fviz\_dend(Clust\_Ave, cex = 0.4, main = "Cluster Dendrogram Average Linkage")

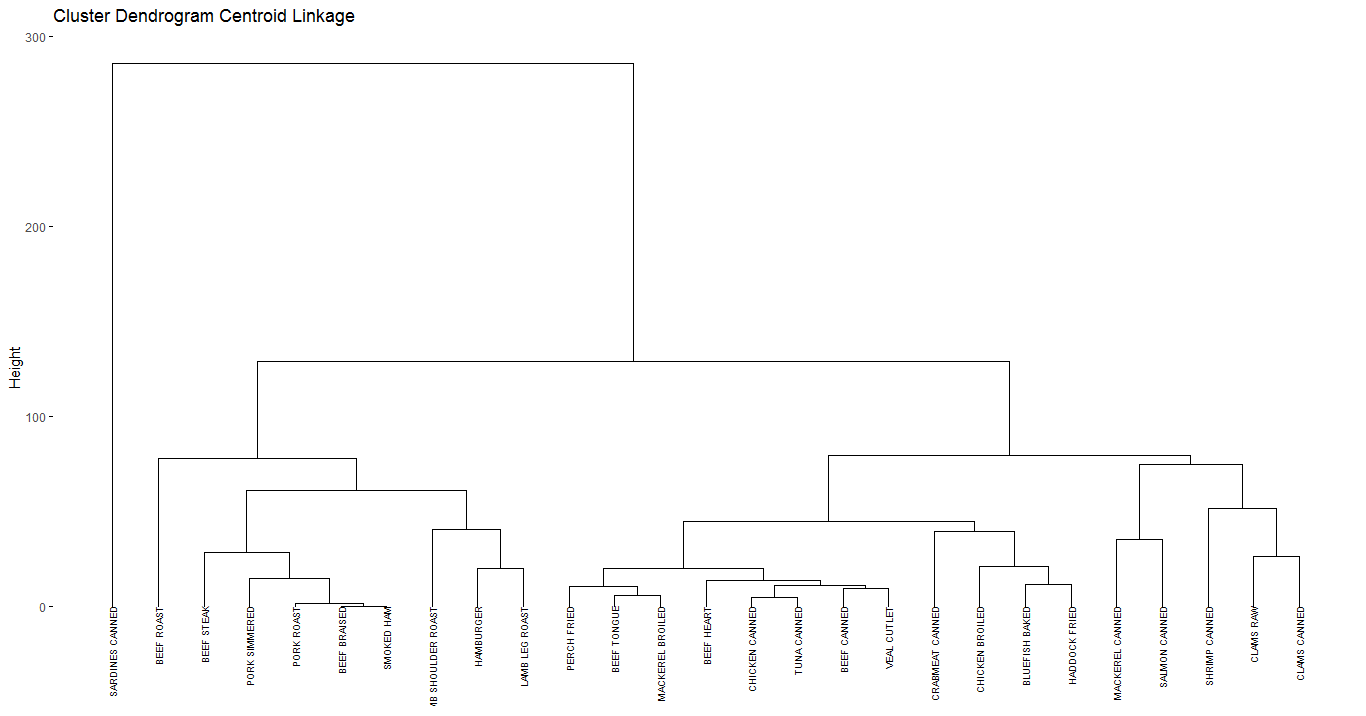


# Centroid Linkage

Clust\_Cen <- hclust(dist(DataClust), method = "centroid")

Clust\_Cen

fviz\_dend(Clust\_Cen, cex = 0.4, main = "Cluster Dendrogram Centroid Linkage")

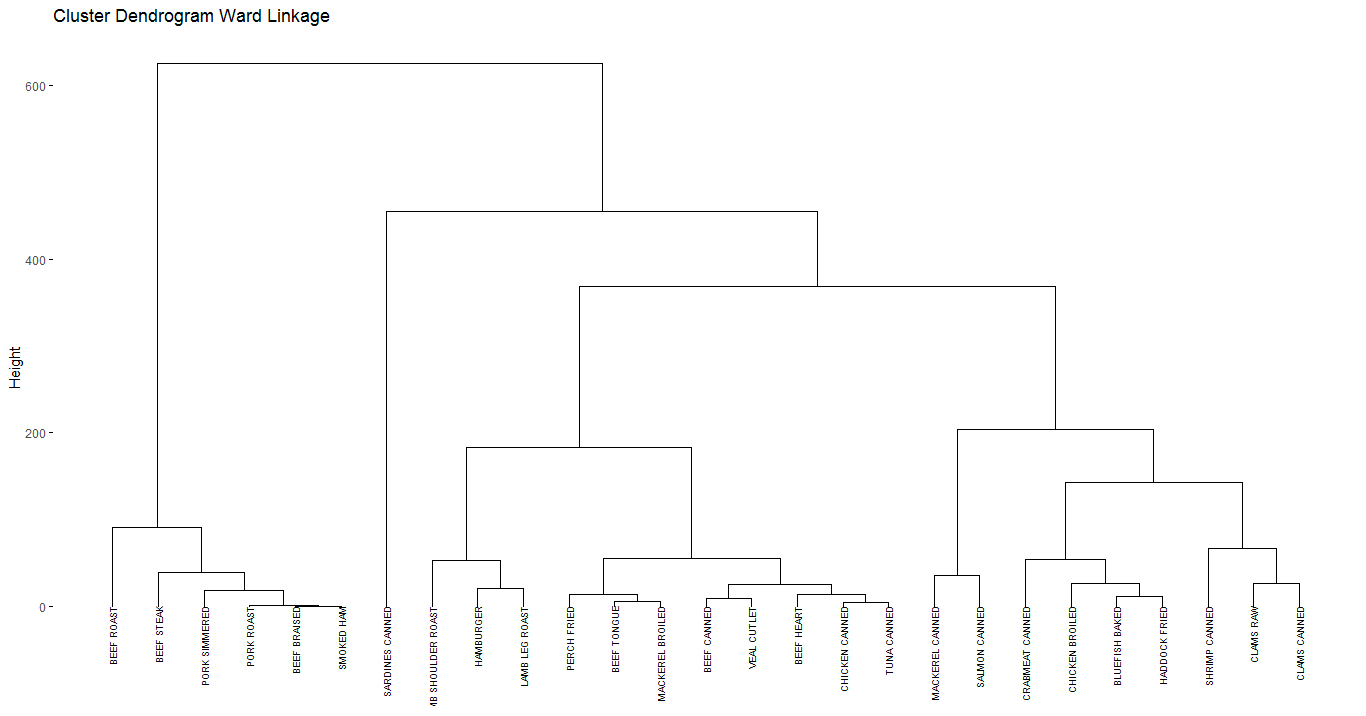


# Ward Linkage

Clust\_War <- hclust(dist(DataClust), method = "ward.D2")

Clust\_War

fviz\_dend(Clust\_War, cex = 0.4, main = "Cluster Dendrogram Ward Linkage")



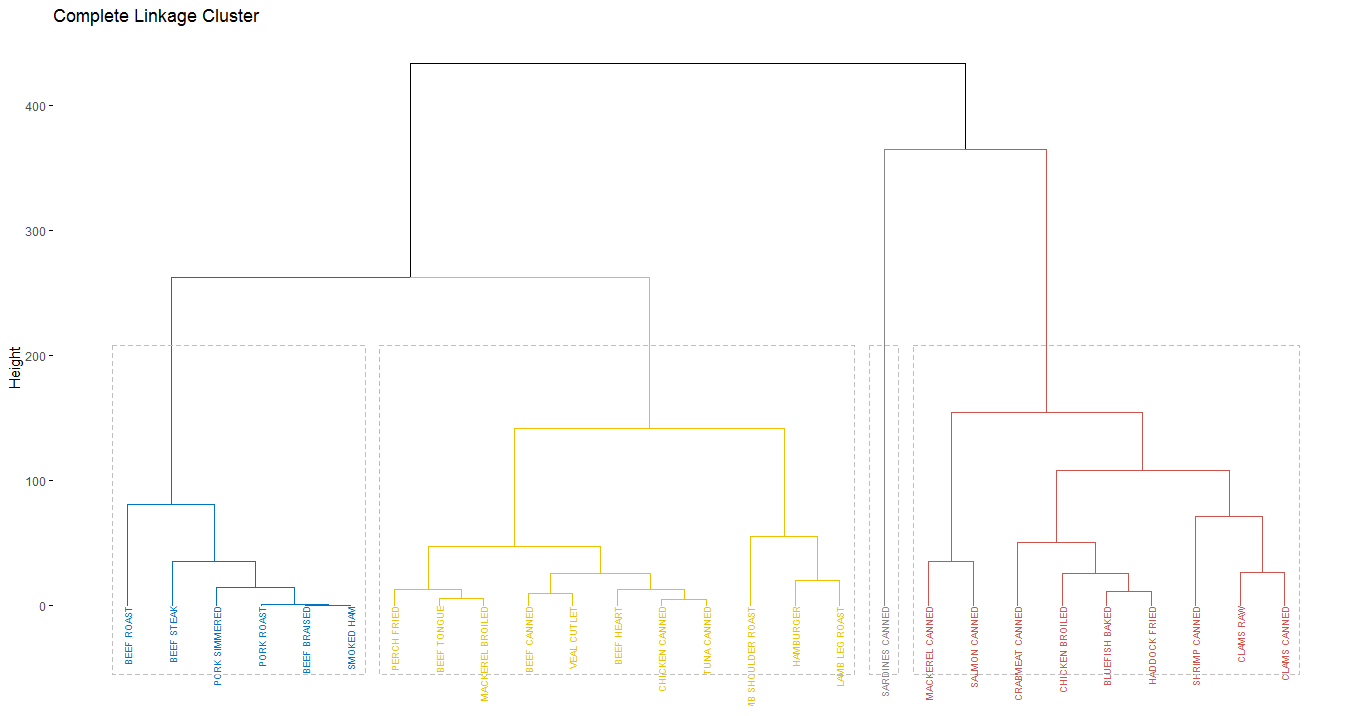
Berdasarkan 5 plot dendogram di atas dengan menggunakan metode aglomerative yang terdiri antara metode centroid, single-linkage, complete-linkage, average-linkage dan ward serta terdapat perhitungan matriks distance dengan fungsi dist(dataset), didapatkan informasi bahwa clustering untuk dataset *nutrient* dapat terbagi menjadi 4 klaster utama. Matriks distance merupakan matriks yang berisikan pasangan jarak antar objek yang dihitung dengan Euclidian Distance.

# Banyaknya Clustering Berdasarkan Syntax dan Dendogram

# Complete Linkage

fviz\_dend(Clust\_Com, k = 4, k\_colors = "jco", rect = TRUE,

cex = 0.4, main = "Complete Linkage Cluster")



Berdasarkan plot dendogram dengan metode complete di atas, data nutrisi dari tiap makanan dapat dikelompokkan menjadi 4 kluster, diantaranya:

* Kluster 1 merupakan kluster dengan jumlah energi (dalam kalori) terbanyak

beef roast, beef steak, pork simmered, pork roast, beef braised, dan smoked ham

* Kluster 2 merupakan kluster dengan jumlah energi menengah ke atas

perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, tuna canned, lamb shoulder roast, hamburger, dan lamb leg roast

* Kluster 3 merupakan kluister dengan jumlah energi menengah ke bawah

sardines canned

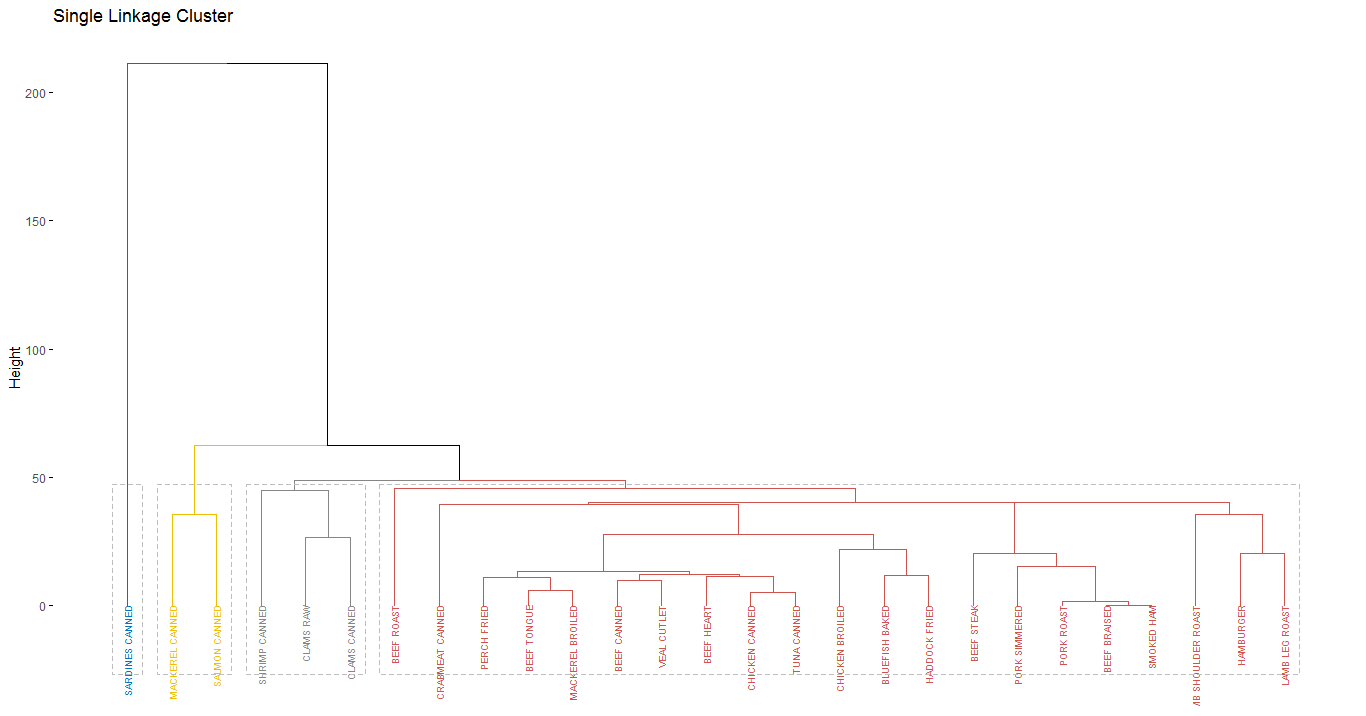
* Kluster 4 merupakan kluster dengan jumlah energy terkecil

mackerel canned, salmon canned, crabmeat canned, chicken broiled, bluefish baked, haddock fried, shrimp canned, clams raw, dan clams canned

# Single Linkage

fviz\_dend(Clust\_Sin, k = 4, k\_colors = "jco", rect = TRUE,

cex = 0.4, main = "Single Linkage Cluster")



Berdasarkan plot dendogram dengan metode single di atas, data nutrisi dari tiap makanan dapat dikelompokkan menjadi 4 kluster, diantaranya:

* Kluster 1 merupakan kluster dengan jumlah kalsium (mg) terbanyak

sardines canned

* Kluster 2 merupakan kluster dengan jumlah kalsium (mg) menengah ke atas

mackerel canned dan salmon canned

* Kluster 3 merupakan kluster dengan jumlah kalsium (mg) menengah ke bawah

shrimp canned, clams raw, dan clams canned

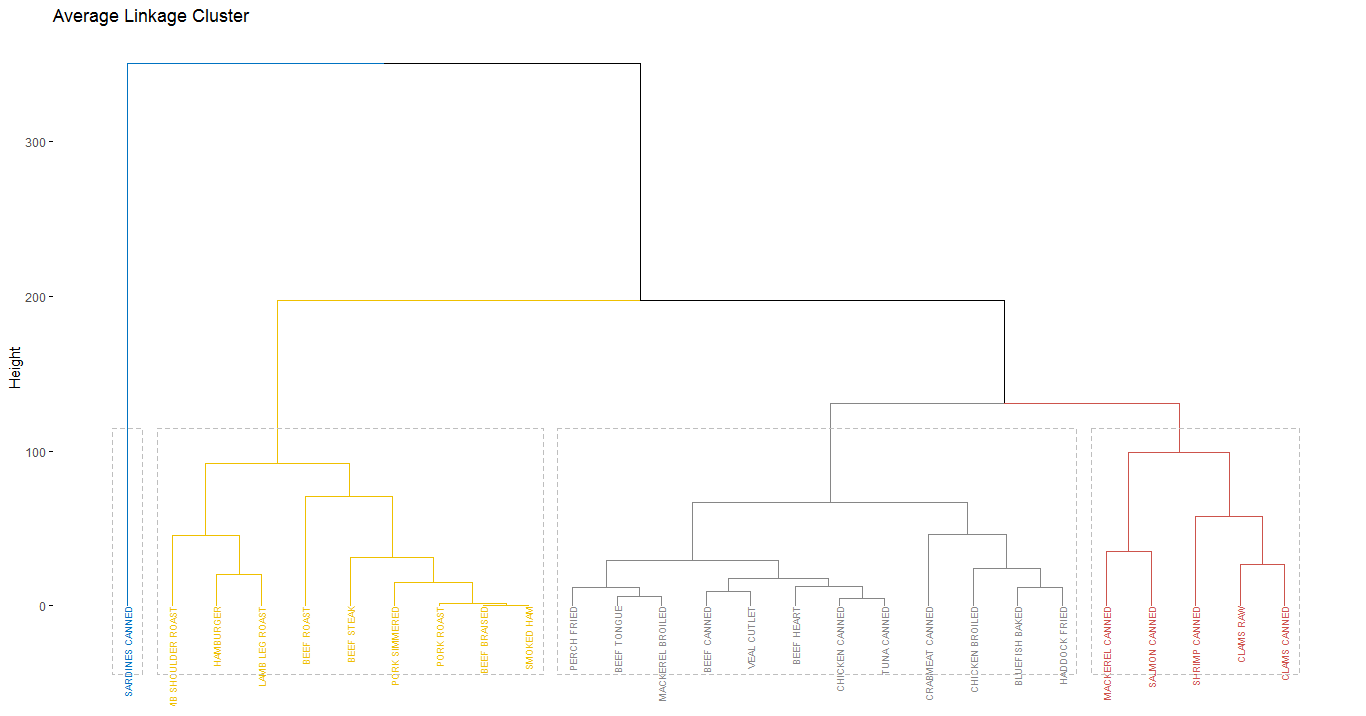
* Kluster 4 merupakan kluster dengan jumlah kalsium (mg) terkecil

beef roast, crabmeat canned, perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, tuna canned, chicken broiled, bluefish baked, haddock fried, pork roast, beef braised, smoked ham, lamb shoulder roast, hamburger, dan lamb leg roast

# Average Linkage

fviz\_dend(Clust\_Ave, k = 4, k\_colors = "jco", rect = TRUE,

cex = 0.4, main = "Average Linkage Cluster")



Berdasarkan plot dendogram dengan metode average di atas, data nutrisi dari tiap makanan dapat dikelompokkan menjadi 4 kluster, diantaranya:

* Kluster 1 merupakan kluster dengan rata-rata nutrisi terbanyak

sardines canned

* Kluster 2 merupakan kluster dengan rata-rata nutrisi menengah ke atas

lamb shoulder roast, hamburger, lamb leg roast, beef roast, beef steak, pork simmered, pork roast, beef braised, dan smoked ham

* Kluster 3 merupakan kluster dengan rata-rata nutrisi menengah ke bawah

perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, tuna canned, crabmeat canned, chicken broiled, bluefish baked, dan haddock fried

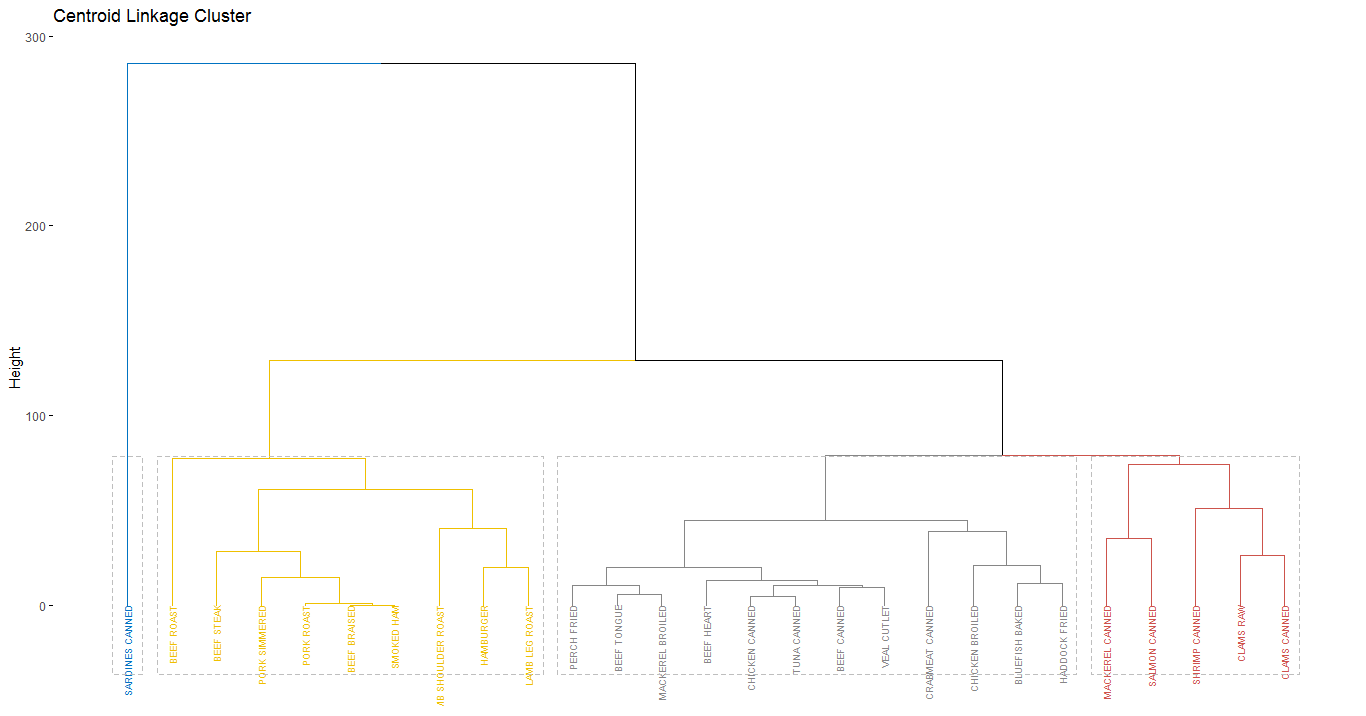
* Kluster 4 merupakan kluster dengan rata-rata nutrisi terkecil

mackerel canned, salmon canned, shrimp canned, clams raw, dan clams canned

# Centroid Linkage

fviz\_dend(Clust\_Cen, k = 4, k\_colors = "jco", rect = TRUE,

cex = 0.4, main = "Centroid Linkage Cluster")



Berdasarkan plot dendogram dengan metode centroid di atas, data nutrisi dari tiap makanan dapat dikelompokkan menjadi 4 kluster, diantaranya:

* Kluster 1 merupakan kluster dengan jarak menuju nilai pusat terdekat

sardines canned

* Kluster 2 merupakan kluster dengan jarak menuju nilai pusat menengah dekat

beef roast, beef steak, pork simmered, pork roast, beef braised, smoked ham, lamb shoulder roast, hamburger, dan lamb leg roast

* Kluster 3 merupakan kluster dengan jarak menuju nilai pusat menengah jauh

perch fried, beef tongue, mackerel broiled, beef heart, chicken canned, tuna canned, beef canned, veal cutlet, crabmeat canned, chicken broiled, bluefish baked, dan haddock fried

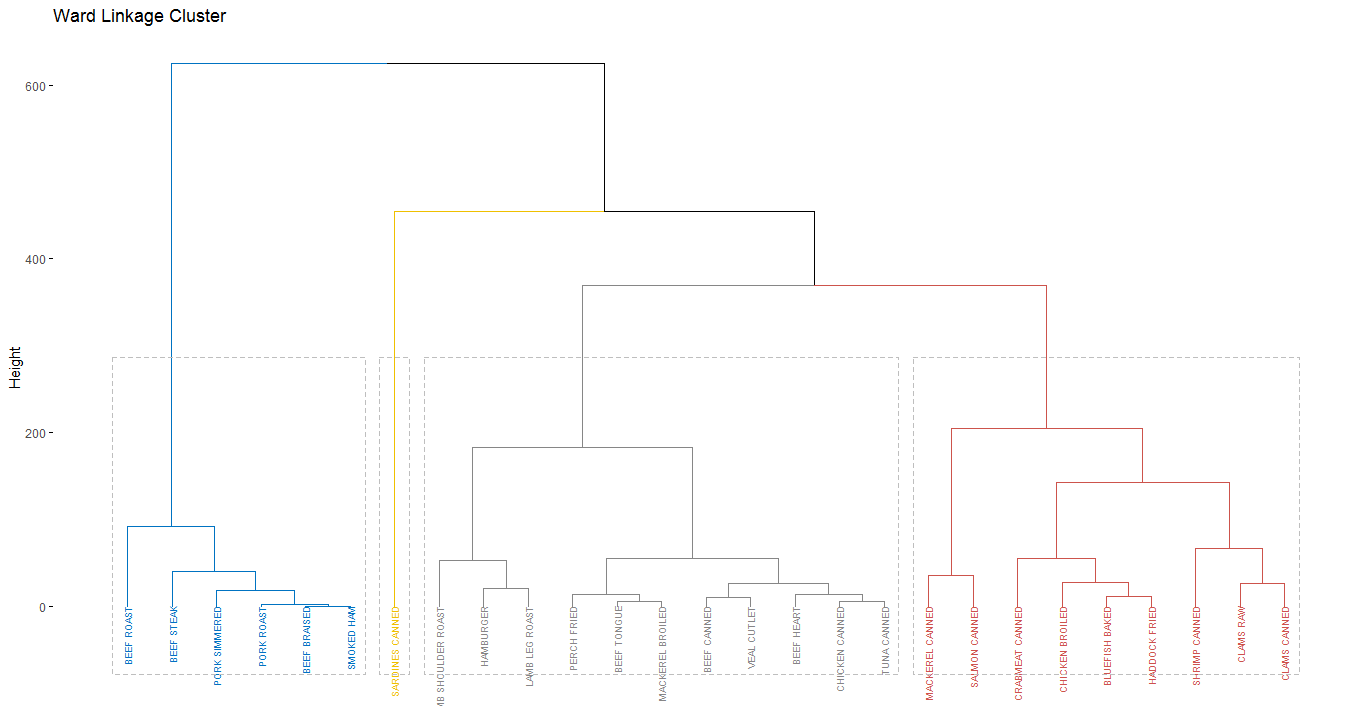
* Kluster 4 merupakan kluster dengan jarak menuju nilai pusat terjauh

mackerel canned, salmon canned, shrimp canned, clams raw, dan clams canned

# Ward Linkage

fviz\_dend(Clust\_War, k = 4, k\_colors = "jco", rect = TRUE,

cex = 0.4, main = "Ward Linkage Cluster")



Berdasarkan plot dendogram dengan metode ward di atas, data nutrisi dari tiap makanan dapat dikelompokkan menjadi 4 kluster, diantaranya:

* Kluster 1 merupakan kluster dengan jumlah beberapa variabel terbanyak

beef roast, beef steak, pork simmered, pork roast, beef braised, dan smoked ham

* Kluster 2 merupakan kluster dengan jumlah beberapa variabel menengah ke atas

sardines canned

* Kluster 3 merupakan kluster dengan jumlah beberapa variabel menengah ke bawah

lamb shoulder roast, hamburger, lamb leg roast, perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, dan tuna canned

* Kluster 4 merupakan kluster dengan jumlah beberapa variabel terkecil

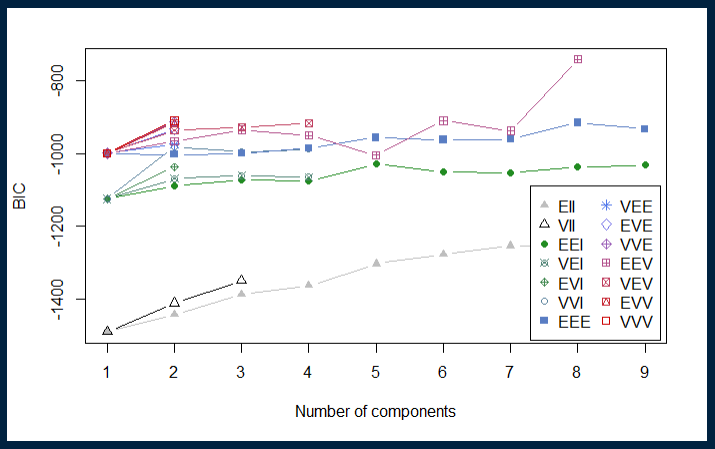
mackerel canned, salmon canned, crabmeat canned, chicken broiled, bluefish baked, haddock fried, shrimp canned, clams raw, dan clams canned

# Banyaknya Clustering Berdasarkan BIC

New\_DataClust <- as.matrix(nutrient)

New\_DataClust <- Mclust(nutrient)

plot(New\_DataClust)

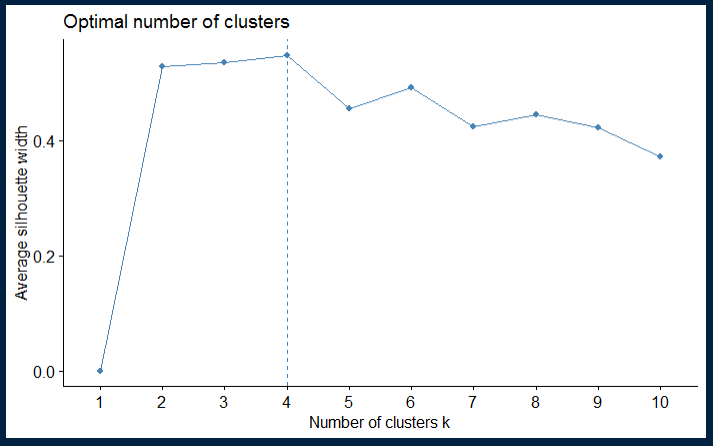


Berdasarkan dari plot BIC tersebut di dapatkan bahwa data *nutrient* dapat dibagi menjadi 1 dan 4. Jika dilihat berdasarkan plot nampak bahwa clustering terbaik yaitu dibagi menjadi 4 kluster. Hal ini ditunjukkan dengan semakin banyak komponen yang ada pada suatu *number of components* maka dapat disimpulkan bahwa *number* tersebut bisa menjadi *clustering* terbaik.

|  |  |  |  |
| --- | --- | --- | --- |
| No | Metode | Nomer klaster | Anggota klaster |
| 1 | Complete-Linkage | 1 | beef roast, beef steak, pork simmered, pork roast, beef braised, dan smoked ham |
| 2 | perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, tuna canned, lamb shoulder roast, hamburger, dan lamb leg roast |
| 3 | sardines canned |
| 4 | mackerel canned, salmon canned, crabmeat canned, chicken broiled, bluefish baked, haddock fried, shrimp canned, clams raw, dan clams canned |
| 2 | Single-Linkage | 1 | sardines canned |
| 2 | mackerel canned dan salmon canned |
| 3 | shrimp canned, clams raw, dan clams canned |
| 4 | beef roast, crabmeat canned, perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, tuna canned, chicken broiled, bluefish baked, haddock fried, pork roast, beef braised, smoked ham, lamb shoulder roast, hamburger, dan lamb leg roast |
| 3 | Average-Linkage | 1 | sardines canned |
| 2 | lamb shoulder roast, hamburger, lamb leg roast, beef roast, beef steak, pork simmered, pork roast, beef braised, dan smoked ham |
| 3 | perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, tuna canned, crabmeat canned, chicken broiled, bluefish baked, dan haddock fried |
| 4 | mackerel canned, salmon canned, shrimp canned, clams raw, dan clams canned |
| 4 | Centroid | 1 | sardines canned |
| 2 | beef roast, beef steak, pork simmered, pork roast, beef braised, smoked ham, lamb shoulder roast, hamburger, dan lamb leg roast |
| 3 | perch fried, beef tongue, mackerel broiled, beef heart, chicken canned, tuna canned, beef canned, veal cutlet, crabmeat canned, chicken broiled, bluefish baked, dan haddock fried |
| 4 | mackerel canned, salmon canned, shrimp canned, clams raw, dan clams canned |
| 5 | Ward | 1 | beef roast, beef steak, pork simmered, pork roast, beef braised, dan smoked ham |
| 2 | sardines canned |
| 3 | lamb shoulder roast, hamburger, lamb leg roast, perch fried, beef tongue, mackerel broiled, beef canned, veal cutlet, beef heart, chicken canned, dan tuna canned |
| 4 | mackerel canned, salmon canned, crabmeat canned, chicken broiled, bluefish baked, haddock fried, shrimp canned, clams raw, dan clams canned |

# Nilai K-Means dan Clustering

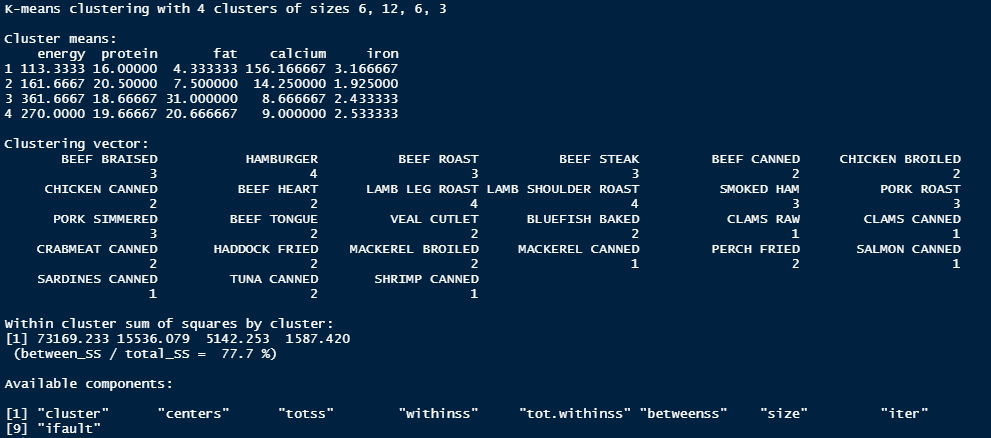
fviz\_nbclust(nutrient, kmeans, method = "silhouette")



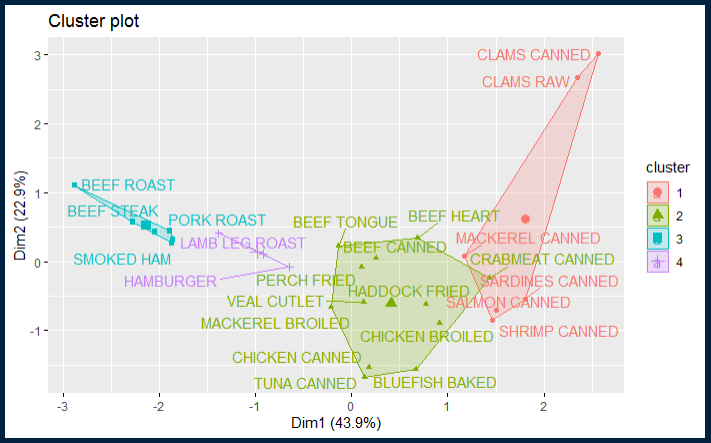
Pada visualisasi dengan metode silhouette, nilai K optimum dapat dilihat dengan titik tertinggi yang ada pada grafik. Namun, bisa juga dilihat dengan titik kedua tertinggi yang ada pada grafik sehingga nilai K optimum pada metode ini adalah antara 3 dan 4. Sehingga dari kedua metode yang digunakan dapat disimpulkan bahwa nilai k optimum adalah 4.

Data\_kmean <- kmeans(nutrient, 2)

Data\_kmean



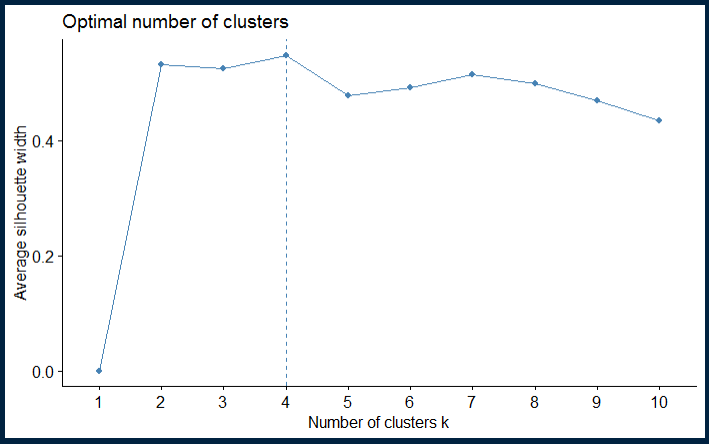
fviz\_cluster(Data\_kmean, data = nutrient, rep = TRUE)



Berdasarkan cluster plot di atas, data *nutrient*  dikelompokkan menjadi 4 kluster. Persebaran data dari dimensi 1 dan 2 sebesar 66,8%. Dari plot tersebut, kluster yang tertinggi adalah kluster 1 dengan data clams canned dan clams raw. Lalu kluster 2 berada di bawahnya. Namun beberapa data pada kluster 2 memiliki nilai dimensi 2 negatif seperti tuna canned dan bluefish baked. Kluster 3 hanya memiliki 2 data yaitu lamb leg roast dan hamburger. Semua data kluster 4 berada pada dimensi 1 negatif.

# Nilai K-Medoids dan Clusteirng

fviz\_nbclust(nutrient, pam, method = "silhouette")



Pada visualisasi dengan metode silhouette, nilai K optimum dapat dilihat dengan titik tertinggi yang ada pada grafik. Namun, bisa juga dilihat dengan titik kedua tertinggi yang ada pada grafik sehingga nilai K optimum pada metode ini adalah antara 3 dan 4. Sehingga dari kedua metode yang digunakan dapat disimpulkan bahwa nilai k optimum adalah 4.

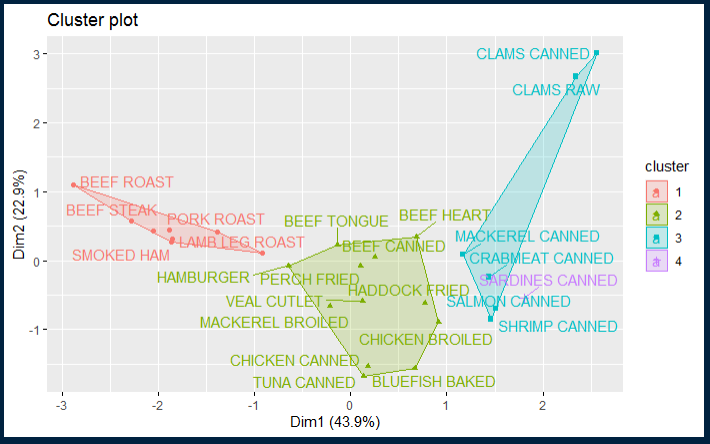
Data\_kmedoid <- pam(nutrient, 2)

summary(Data\_kmedoid)

Data\_kmedoid



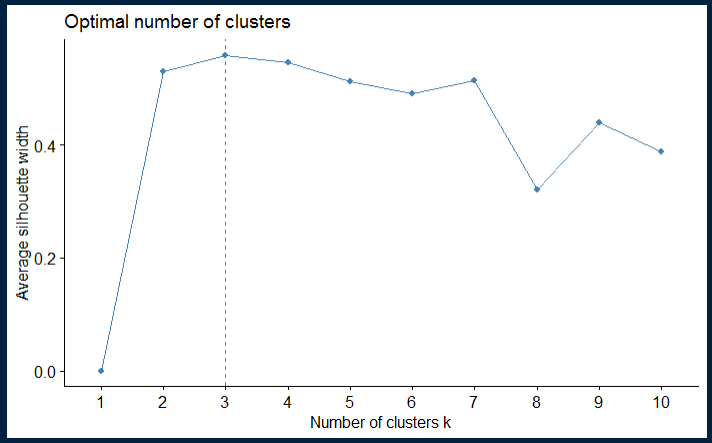
fviz\_cluster(Data\_kmedoid, data = nutrient, rep = TRUE)



Berdasarkan cluster plot di atas, data *nutrient*  dikelompokkan menjadi 4 kluster. Persebaran data dari dimensi 1 dan 2 sebesar 66,8%. Dari plot tersebut, kluster yang tertinggi adalah kluster 3 dengan data clams canned dan clams raw. Lalu kluster 2 berada di bawahnya. Namun beberapa data pada kluster 2 memiliki nilai dimensi 2 negatif seperti tuna canned dan bluefish baked. Kluster 1 hanya memiliki 1 data yaitu sardines caneed dengan nilai dimensi 2 negatif. Semua data kluster 4 berada pada dimensi 1 negatif.

# Nilai KMedians

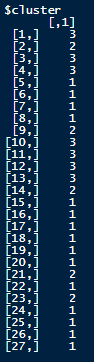
fviz\_nbclust(nutrient, kGmedian, method = "silhouette")

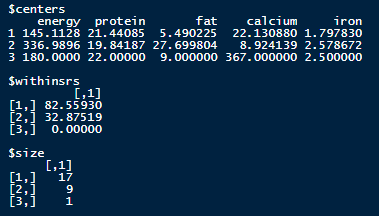


Pada visualisasi dengan metode silhouette, nilai K optimum dapat dilihat dengan titik tertinggi yang ada pada grafik. Namun, bisa juga dilihat dengan titik kedua tertinggi yang ada pada grafik sehingga nilai K optimum pada metode ini adalah antara 2, 3, dan 4. Sehingga dari kedua metode yang digunakan dapat disimpulkan bahwa nilai k optimum adalah 3.

Data\_kmedian <- kGmedian(nutrient, 3)

Data\_kmedian

```



**Kesimpulan**

Dataset *nutrient* sangat optimal apabila dikelompokkan menjadi 4 kluster, baik menggunakan metode complete-linkage, single-linkage, average-linkage, centroid, maupun ward. Masing-masing metode memiliki pembagian data yang belum tentu sama dengan metode yang lain.