Analisis Gerombol

gdito

Note: output dari R pada dokumen ini diawali dengan tanda ##

Package

Pada Praktikum kali ini package yang dibutuhkan adalah

- factoextra

Silahkan install jika belum ada

```
install.packages("factoextra")
```

```
library("factoextra")
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

Metode K-means

Prosedur penerapan K-means

1. Pre-processing data
2. Memilih banyaknya gerombol
3. Menerapkan K-means
4. Interprestasi Gerombol yang terbentuk

Data Pelanggan Mall

Seorang pemilik Mall ingin mengelompokan customer di Mall yang ia miliki, sehingga tim marketing bisa mengembangkan strategi yang tepat untuk customer yang tepat pula. Data yang dimiliki oleh Mall tersebut adalah Customer ID, umur pelanggan (age), pendapatan tahunan dalam ribu dollar (annual income) dan spending score. Spending score merupakan nilai yang diberikan oleh Mall kepada customer berbasarkan perilaku customer (waktu kunjungan,jenis barang yang dibeli, dan banyaknya uang yang dihabiskan dalam belanja) yang memiliki rentang nilai 1-100. Semakin besar nilai Spending Score berarti customer semakin loyal pada Mall tersebut dan semakin besar pula uang belanja yang digunakan.

Menyiapkan data di R

```
data_mall <- read.csv("Mall_Customers.csv")
head(data_mall)
```

```
##   CustomerID  Genre Age Annual.Income Spending.Score
## 1          1   Male  19            15             39
## 2          2   Male  21            15             81
## 3          3 Female  20            16              6
## 4          4 Female  23            16             77
## 5          5 Female  31            17             40
## 6          6 Female  22            17             76
```

Pre-processing data

Peubah yang digunakan untuk menerapkan k-means adalah peubah Age AnnualIncome dan Spending Score. Oleh karena itu peubah yang tidak kita gunakan akan kita hilangkan tersebih dahulu.

```
data_mall <- data_mall[,c("Age","Annual.Income","Spending.Score")]
head(data_mall)
```

```
##   Age Annual.Income Spending.Score
## 1  19            15             39
## 2  21            15             81
## 3  20            16              6
## 4  23            16             77
## 5  31            17             40
## 6  22            17             76
```

Standarisasi peubah

Standarisasi peubah merupakan proses transformasi peubah menjadi peubah yang memiliki rata-rata nol dan simpangan baku satu. Process standarisasi ini dilakukan jika kita melihat perbedanan satuan pengukuran peubah-peubah yang digunakan contoh(umur dan pendapatan). Standarisasi dilakukan karena metode k-means menggunakan konsep jarak antara objek/amatan, yang mana sensitif terhadap satuan pengukuran. Formula untuk standarisasi data adalah sebagai berikut:

$$y = \frac{y - \bar{y}}{\sigma_y}$$

dengan $\bar{y}$ merupakan rata-rata dari y dan σ_y merupakan simpangan baku dari y.

Dalam R, standarisasi data bisa dilakukan dengan menggunakan fungsi `scale`.

```
data_mall_standardize <- scale(data_mall)
apply(data_mall_standardize,2,mean)
```

```
##            Age  Annual.Income Spending.Score
##  -1.016906e-16  -8.144310e-17  -1.096708e-16
```

```
apply(data_mall_standardize,2,sd)
```

```
##            Age  Annual.Income Spending.Score
##              1              1              1
```

Jika kita perhatikan rata-rata dan simpangan baku peubah setelah distandarisasi mendekati nol dan satu.

Note: Dalam tahapan pre-processing data, kita menyiapkan data agar metode kmeans bisa diterapkan secara maksimal. Dua hal yang umumnya dilakukan pada tahap ini adalah memilih peubah yang digunakan dan melakukan standarisasi peubah.

Memilih banyaknya gerombol

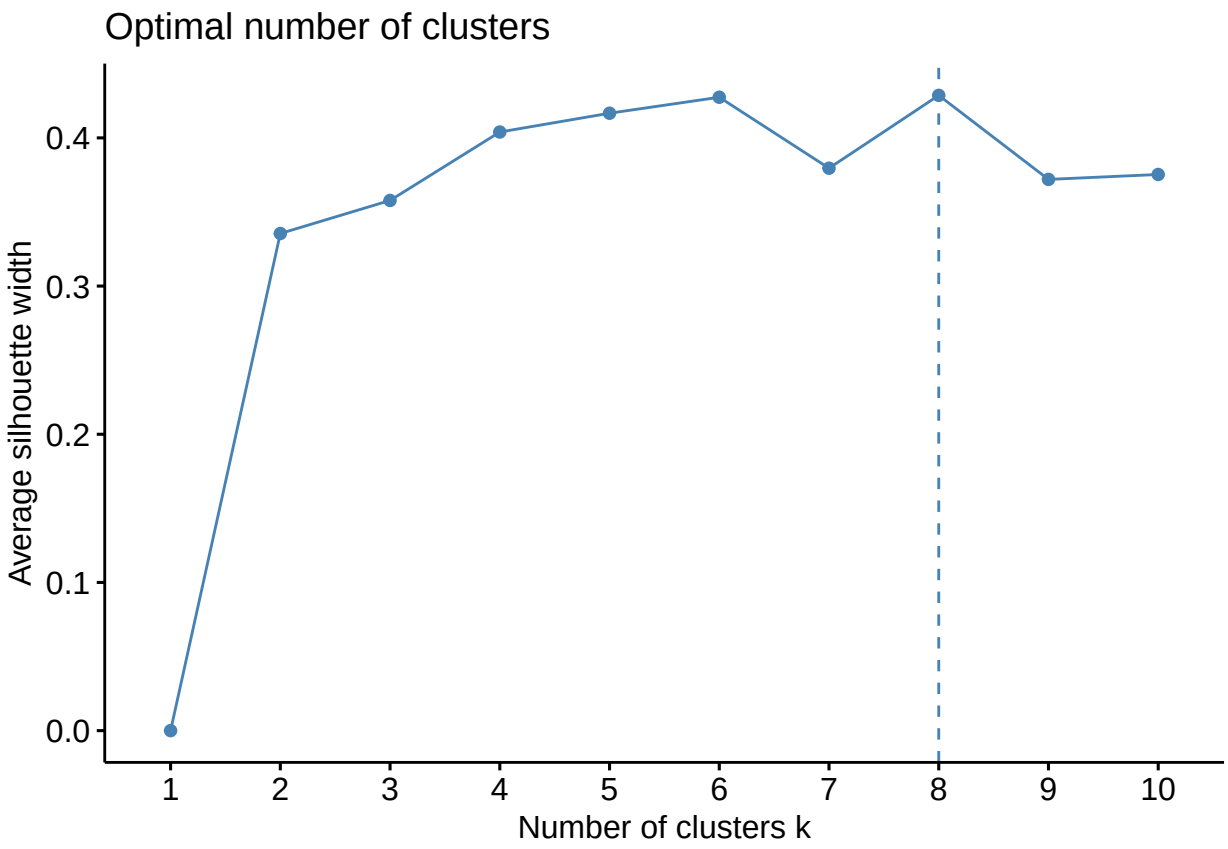
Umumnya, banyaknya gerombol dapat ditentukan dengan menggunakan beberapa kriteria statistik, seperti koefisien **silhouette** dan **WSS** atau (Within Sum of Square).

Kriteria koefisien silhouette dihitung berdasarkan jarak antar amatan. Koefisien ini mengukur seberapa dekat suatu amatan dengan amatan lain yang berada di gerombol yang sama (dikenal sebagai ukuran cohesion) dibandingkan dengan jarak terhadap amatan lain yang berada di gerombol berbeda (dikenal sebagai ukuran separation). Koefisien yang nilainya semakin besar menunjukkan bahwa gerombol yang terbentuk sudah sesuai.

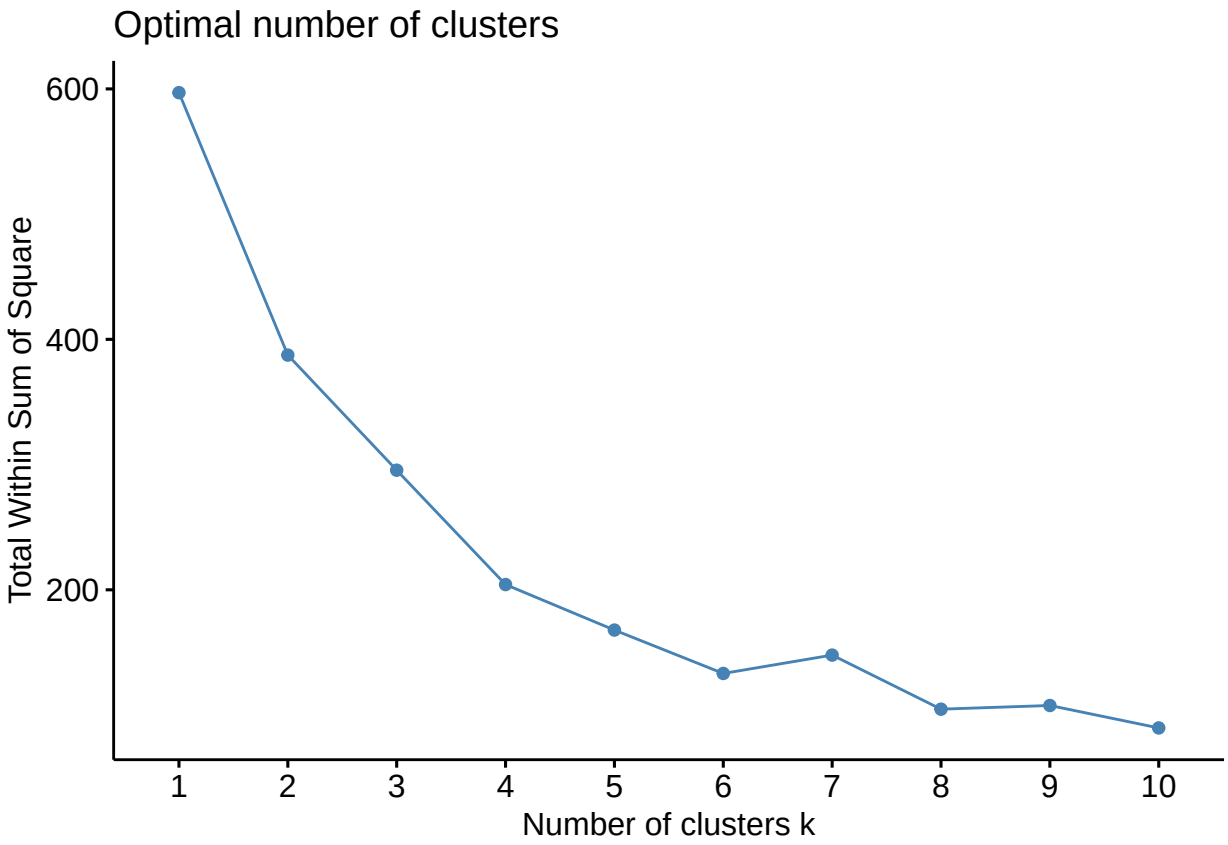
Kriteria WSS merupakan kriteria yangmenghitung keragamaan dalam gerombol yang terbentuk. Semakin kecil keragaman dalam gerombol yang terbentuk menunjukkan bahwa gerombol yang terbentuk sudah sesuai.

Dengan menggunakan kriteria tersebut, kita bisa membandingkan banyaknya gerombol yang paling sesuai pada data yang kita sedang analisis. Dalam R, fungsi `fviz_nbclust` dari package `factoextra` dapat digunakan untuk memilih banyaknya gerombol.

```
fviz_nbclust(data_mall_standardize,FUNcluster = kmeans,method = "silhouette")
```




```
fviz_nbclust(data_mall_standardize,FUNcluster = kmeans,method = "wss")
```



Untuk kriteria koefisien silhoutte, banyaknya gerombol dengan nilai koefisien tertinggi yang kita pilih. Sedangkan pada WSS, banyaknya gerombol yang kita pilih didasarkan pada banyaknya gerombol yang mana garisnya berbentuk seperti siku (elbow). Pada gambar diatas garis membentuk siku saat berada di gerombol keempat. **Karena penentuan ini berdasarkan visual,jadi setiap orang mungkin berbeda melihat pola sikunya**

Berdasarkan kedua kriteria tersebut, banyaknya gerombol terbaik yang dipilih berbeda. Jika demikian, banyaknya gerombol bisa ditentukan berdasarkan kemudahan interpretasi gerombol yang terbentuk. Pada tulisan ini kita akan menggunakan 4 gerombol saja.

Note: secara default banyaknya gerombol yang dicobakan pada fungsi `fviz_nbclust` adalah 10, jika ingin merubah hal tersebut bisa dilakukan dengan menggunakan argumen `kmax` dalam fungsi,misal `kmax=20`.

Menerapkan K-means

Setelah kita mendapatkan banyaknya gerombol terbaik, maka selajutnya kita akan menerapkan metode kmenas untuk mendapatkan label gerombol pada setiap amatan. Fungsi `eclust` dari package `factoextra` digunakan untuk menerpkan metode kmeans. Pada fungsi `eclust`, kita cukup memasukan data yang sebelum distandarisasi, karena dalam fungsi tersebut terdapat argumen `stand`, yang jika diatur `stand=TRUE` secara otomatis data yang kita gunakan akan distandarisasi.

```
kmeans_mall <- eclust(data_mall,stand = TRUE,FUNcluster = "kmeans",k=4,graph = F)
kmeans_mall$cluster
```

```
##   [1] 3 3 3 3 3 3 2 3 2 3 2 3 2 3 2 3 3 3 2 3 3 3 2 3 2 3 2 3 2 3 2 3 2 3 2
##  [36] 3 2 3 2 3 2 3 2 3 2 3 2 3 3 3 2 3 3 2 2 2 2 2 3 2 2 3 2 2 2 3 2 2 3 3
##  [71] 2 2 2 2 2 3 2 2 3 2 2 3 2 2 3 2 2 3 3 2 2 3 2 2 3 3 2 3 2 3 3 2 2 3 2
## [106] 3 2 2 2 2 2 3 1 3 3 3 2 2 2 2 3 1 4 4 1 4 1 4 2 4 1 4 1 4 1 4 1 4 1 4
```

```
## [141] 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 2 4 1 4 1 4 1 4 1 4 1 4 1 4 1
## [176] 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4
```

```
kmeans_mall$centers
```

```
##           Age Annual.Income Spending.Score
## 1  0.03711223     0.9876366     -1.1857814
## 2  1.08344244    -0.4893373     -0.3961802
## 3 -0.96008279    -0.7827991      0.3910484
## 4 -0.42773261     0.9724070      1.2130414
```

Label gerombol untuk setiap amatan/objek, bisa diperoleh dengan menggunakan `$cluster`. Kemudian,interpretasi setiap gerombol yang terbentuk dapat dilakukan dengan menggunakan bantuan nilai rata-rata dari masing-masing peubah dihitung berdasarkan gerombol. Informasi ini bisa diperoleh dengan menggunakan `$centers`. Karena kita melakukan standarisasi peubah, maka nilai rata-rata yang diperoleh juga dalam skala standarisasi.

4. Interprestasi Gerombol yang terbentuk

Berdasarkan nilai rata-rata dari `$centers`, berikut adalah interpretasinya

- Gerombol 1 : gerombol ini merupakan customer-customer yang cukup muda (peubah age bernilai kecil) dan berpenghasilan besar (peubah Income bernilai besar) namun sedikit sekali menghabiskan uangnya untuk berbelanja (peubah spending score bernilai kecil bahkan negatif).
- Gerombol 2 : gerombol ini merupakan customer-customer yang sudah tua (peubah age bernilai besar) dan berpenghasilan kecil (peubah Income bernilai kecil) dan sedikit sekali menghabiskan uangnya untuk berbelanja (peubah spending score bernilai kecil). Gerombol ini mungkin murupakan customer yang sudah pensiun dan hanya memiliki pemasukan dari tunjangan pensiun.
- Gerombol 3 : gerombol ini merupakan customer-customer yang masih sangat muda (peubah age bernilai kecil) dan berpenghasilan kecil (peubah Income bernilai kecil) namun menghabiskan uangnya untuk berbelanja cukup besar (peubah spending score bernilai besar). Gerombol ini mungkin murupakan customer yang aneh, karena memiliki penghasilan yang kecil namun belanjanya banyak.
- Gerombol 4 : gerombol ini merupakan customer-customer yang masih cukup muda (peubah age bernilai kecil) dan berpenghasilan besar (peubah Income bernilai besar) namun menghabiskan uangnya untuk berbelanja cukup besar (peubah spending score bernilai besar). Gerombol ini mungkin murupakan customer yang paling menarik untuk menjadi target marketing selanjutnya.

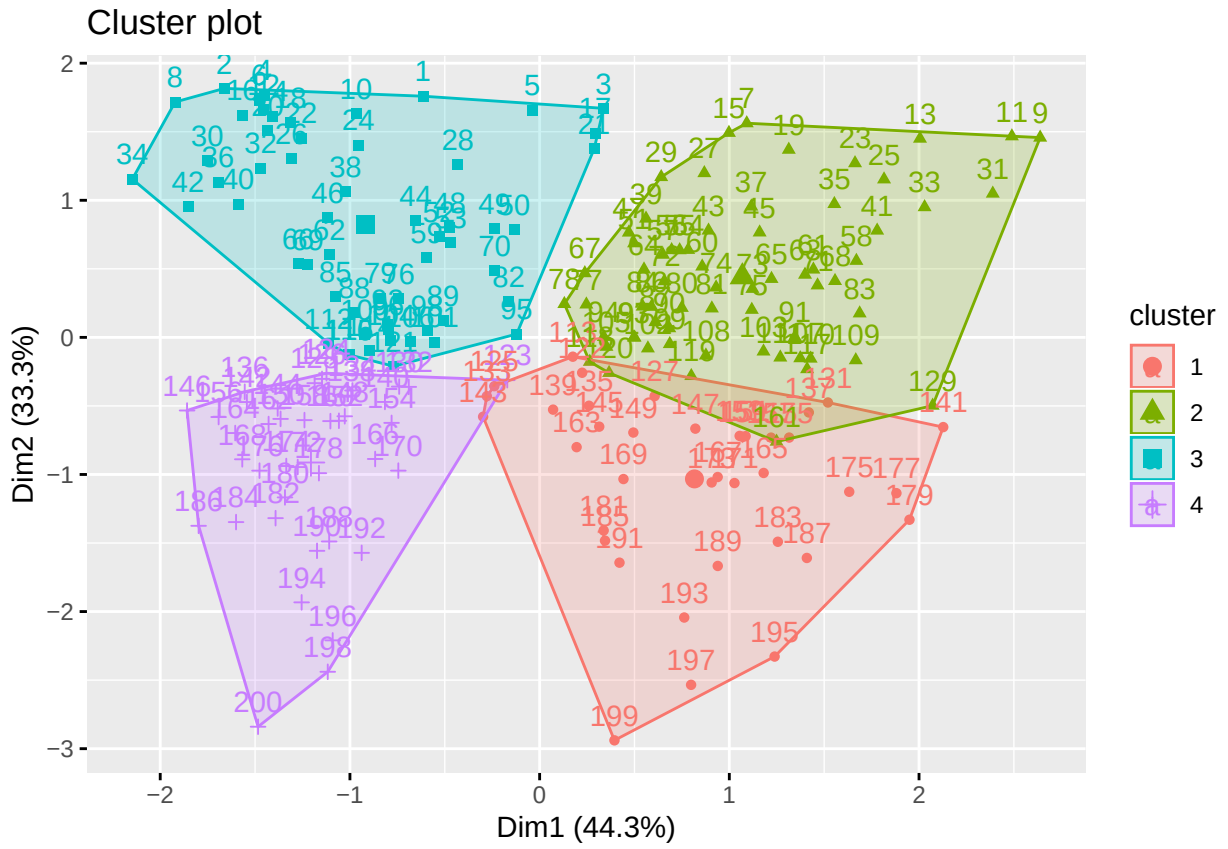
Jika sulit membaca hasil dalam bentuk skala standarisasi maka kita bisa menggunakan fungsi `aggregate` untuk melihat rata-ratanya dalam skala aslinya. Fungsi ini dapat menghitung rata-rata setiap peubah berdasarkan gerombol yang terbentuk.

```
aggregate(data_mall,by =list(gerombol=kmeans_mall$cluster),
            FUN = mean)
```

```
##   gerombol      Age Annual.Income Spending.Score
## 1        1 39.36842      86.50000       19.57895
## 2        2 53.98462      47.70769       39.96923
## 3        3 25.43860      40.00000       60.29825
## 4        4 32.87500      86.10000       81.52500
```

Cara lain untuk memnginterpretasikan hasil gerombol adalah menggunakan scatterplot. Jika peubah untuk membangun kluster lebih dari dua, maka sebelum dibentuk scatterplot peubah tersebut direduksi terlebih dahulu menggunakan analisis komponen utama menjadi dua komponen utama. Namun, untuk interpretasinya setiap gerombolnya kita harus mengetahui interpretasi dari kedua komponen utama dan belum tentu dengan dua komponen utama tersebut sudah mampu menjelaskan keragaman data asal dengan baik.

```
fviz_cluster(kmeans_mall)
```



Interpretasi dua komponen utama bisa dilihat dengan akar cirinya.

```
pca_mall <- prcomp(data_mall_standardize)
pca_mall$rotation
```

```
##                       PC1         PC2         PC3
## Age             0.70638235 -0.03014116 0.707188441
## Annual.Income  -0.04802398 -0.99883160 0.005397916
## Spending.Score -0.70619946  0.03777499 0.707004506
```

Metode Hierarchical Clustering

Prosedur Hierarchical Clustering

1. Pre-processing data
2. Memilih metode linkage dan banyaknya gerombol
3. Menerapkan Hierarchical Clustering
4. Interprestasi Gerombol yang terbentuk

Data yang digunakan untuk ilustrasi Hierarchical Clustering sama seperti Kmeans diatas, yaitu menggunakan data pelanggan Mall

Menyiapkan data di R

```
data_mall <- read.csv("Mall_Customers.csv")
head(data_mall)
```

```
##   CustomerID  Genre Age Annual.Income Spending.Score
## 1          1   Male  19            15             39
## 2          2   Male  21            15             81
## 3          3 Female  20            16              6
## 4          4 Female  23            16             77
## 5          5 Female  31            17             40
## 6          6 Female  22            17             76
```

Pre-processing data

Memilih peubah yang digunakan untuk analisis

```
data_mall <- data_mall[,c("Age","Annual.Income","Spending.Score")]
head(data_mall)
```

```
##   Age Annual.Income Spending.Score
## 1  19            15             39
## 2  21            15             81
## 3  20            16              6
## 4  23            16             77
## 5  31            17             40
## 6  22            17             76
```

Standarisasi peubah

```
data_mall_standardize <- scale(data_mall)
```

Memilih metode linkage dan banyaknya gerombol

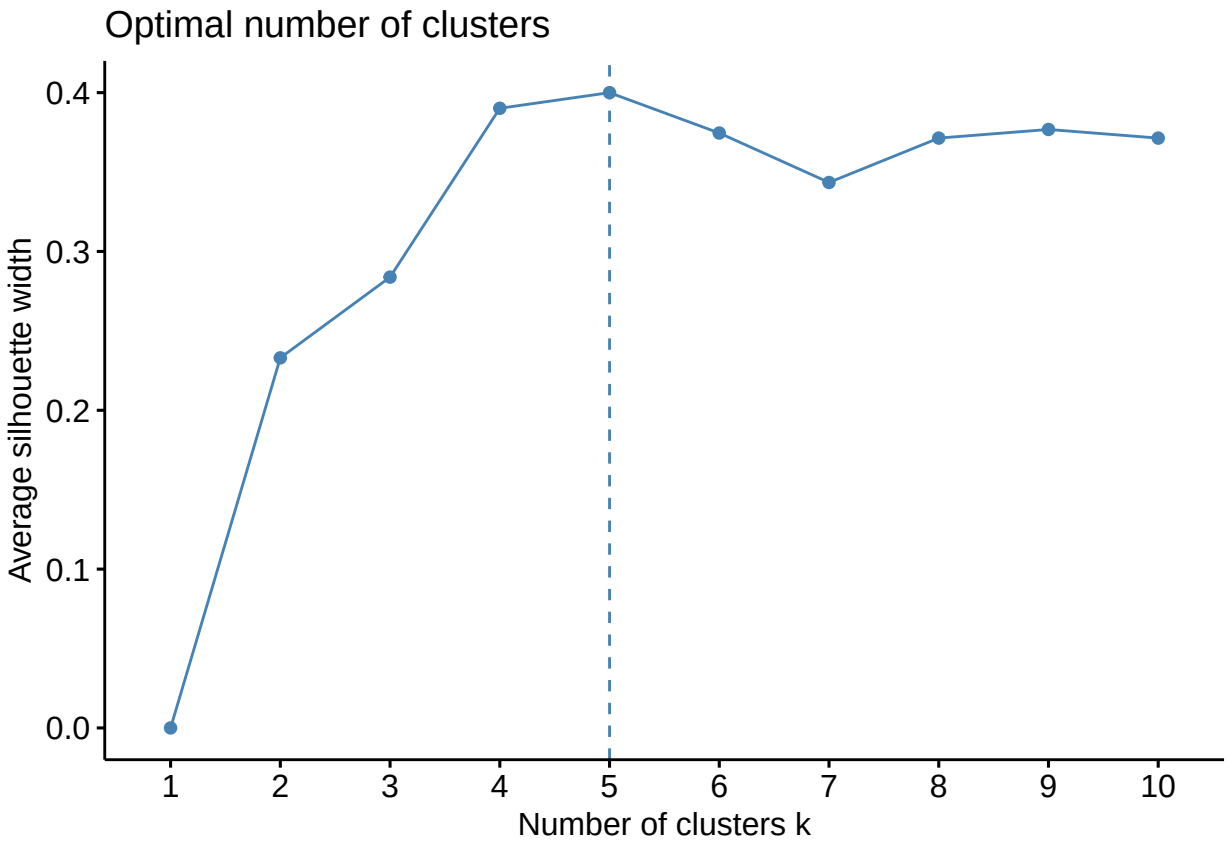
Untuk memilih metode linkage dan banyaknya gerombol bisa menggunakan

a. Koefisien silhoutte dan WSS (seperti k-means)
b. Menggunakan dendogram

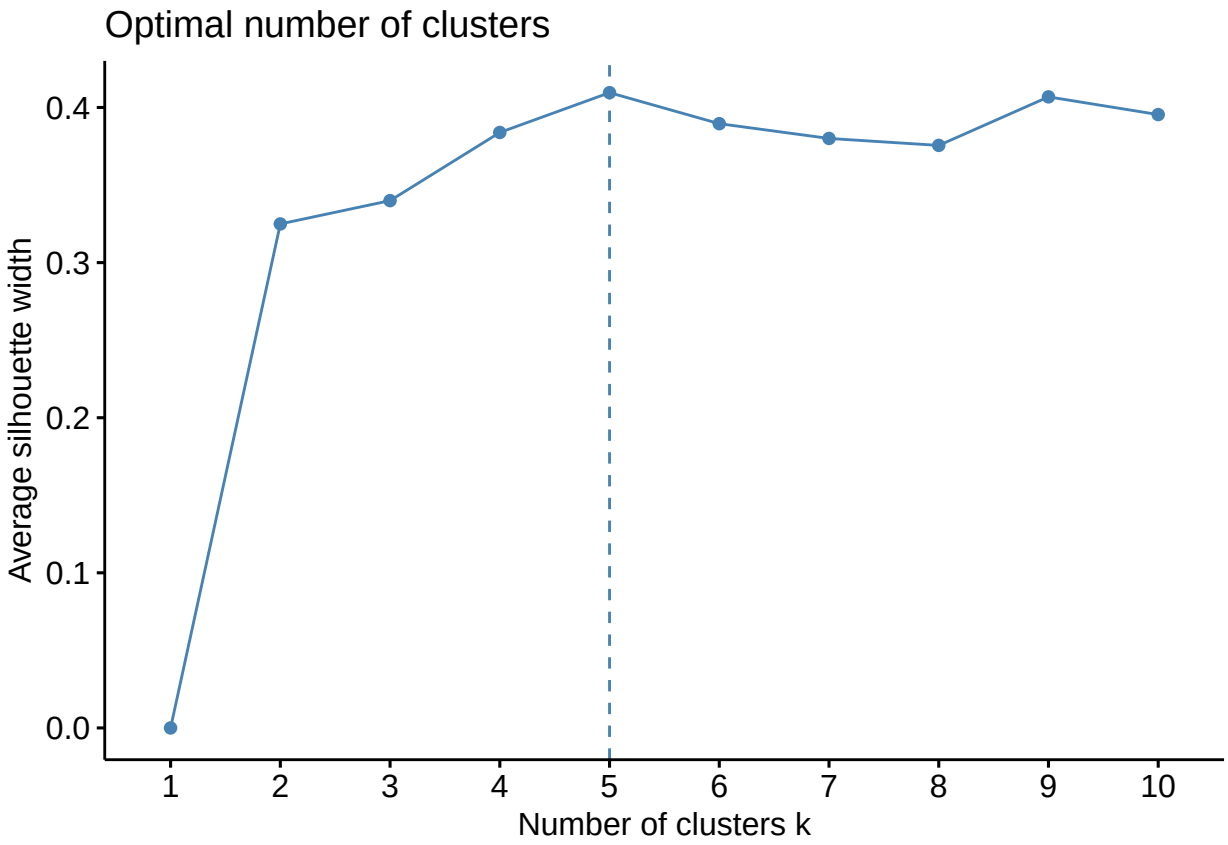
Menggunakan koefisien silhouette dan wss

Untuk ilustrasi kita akan menggunakan metode silhouette saja karena lebih mudah menentukan jumlah gerombolnya.

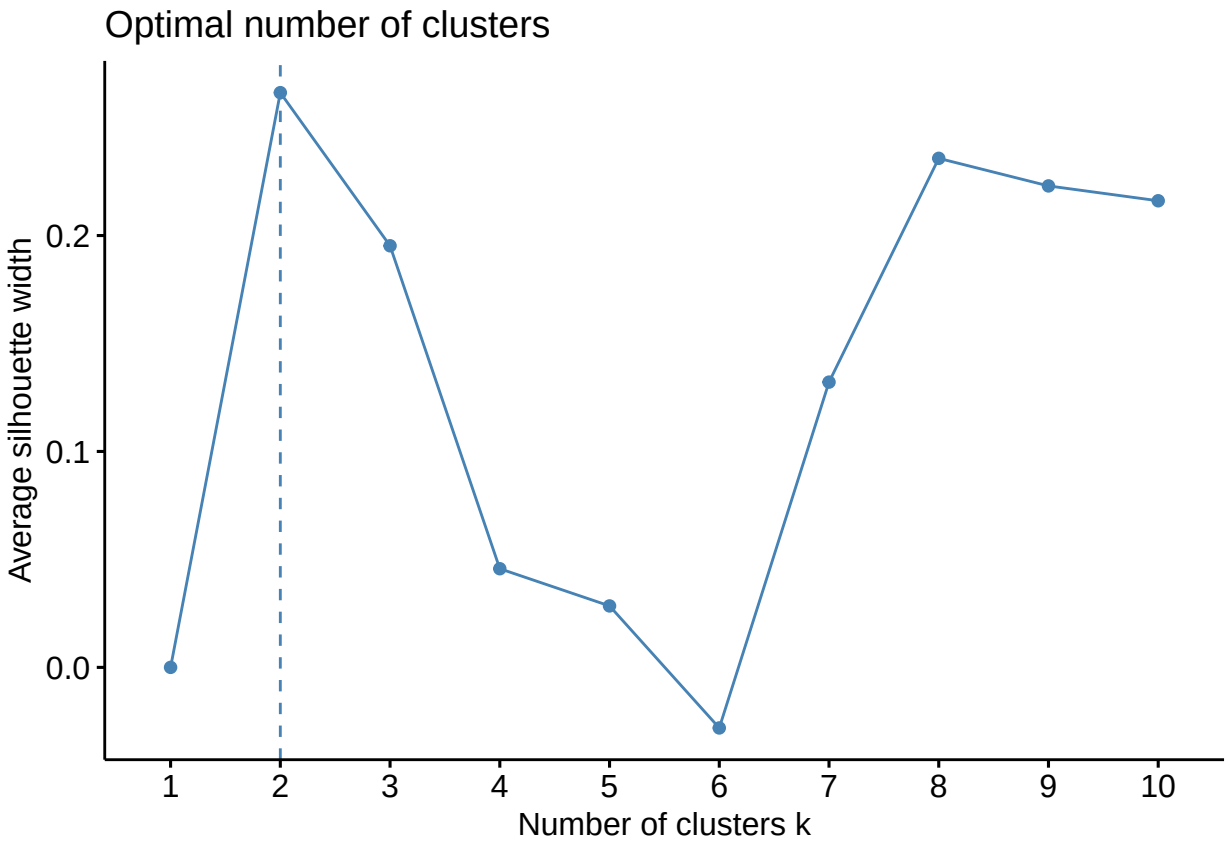
```
#complete
fviz_nbclust(data_mall_standardize,FUNcluster = hcut,method = "silhouette",
             hc_method = "complete",hc_metric = "euclidean")
```



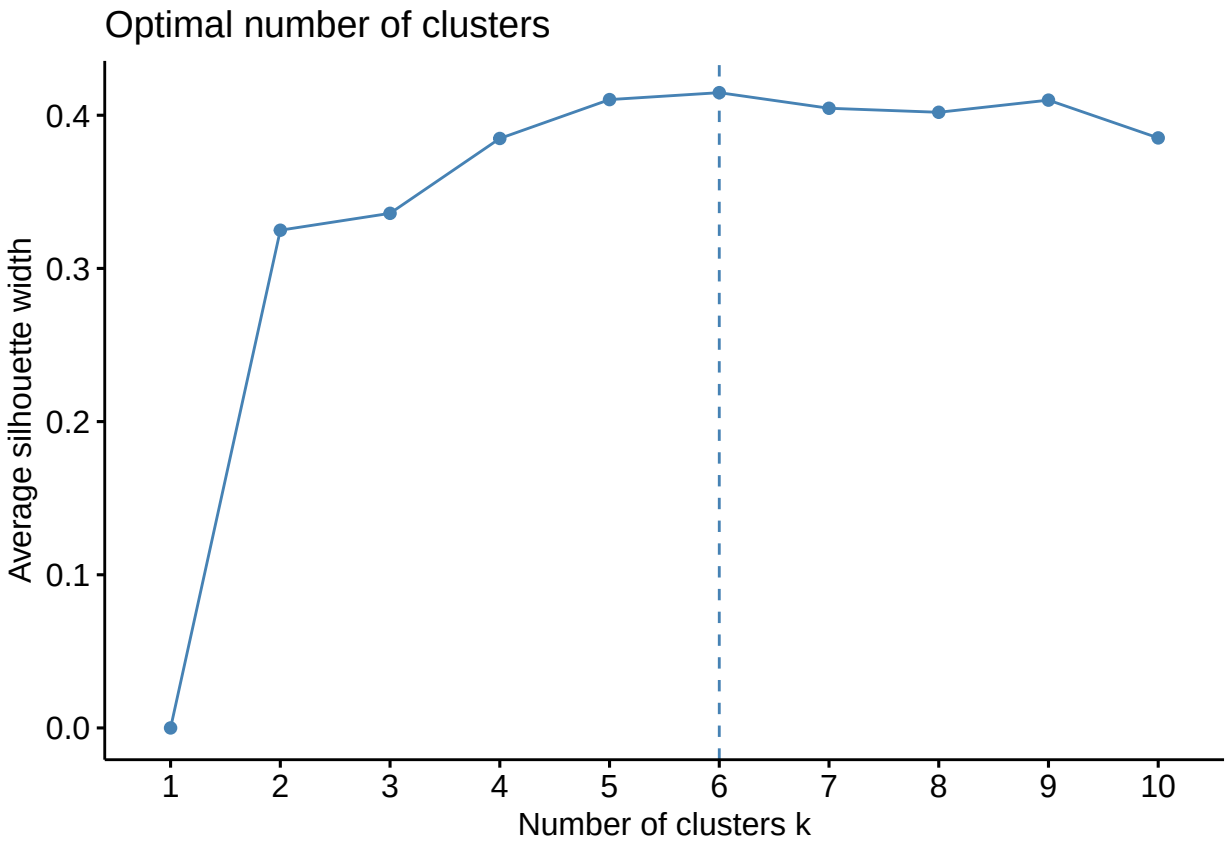

```
#average
fviz_nbclust(data_mall_standardize,FUNcluster = hcut,method = "silhouette",
             hc_method = "average",hc_metric = "euclidean")
```




```
#centroid
fviz_nbclust(data_mall_standardize,FUNcluster = hcut,method = "silhouette",
             hc_method = "centroid",hc_metric = "euclidean")
```




```
#ward
fviz_nbclust(data_mall_standardize,FUNcluster = hcut,method = "silhouette",
             hc_method = "ward.D",hc_metric = "euclidean")
```

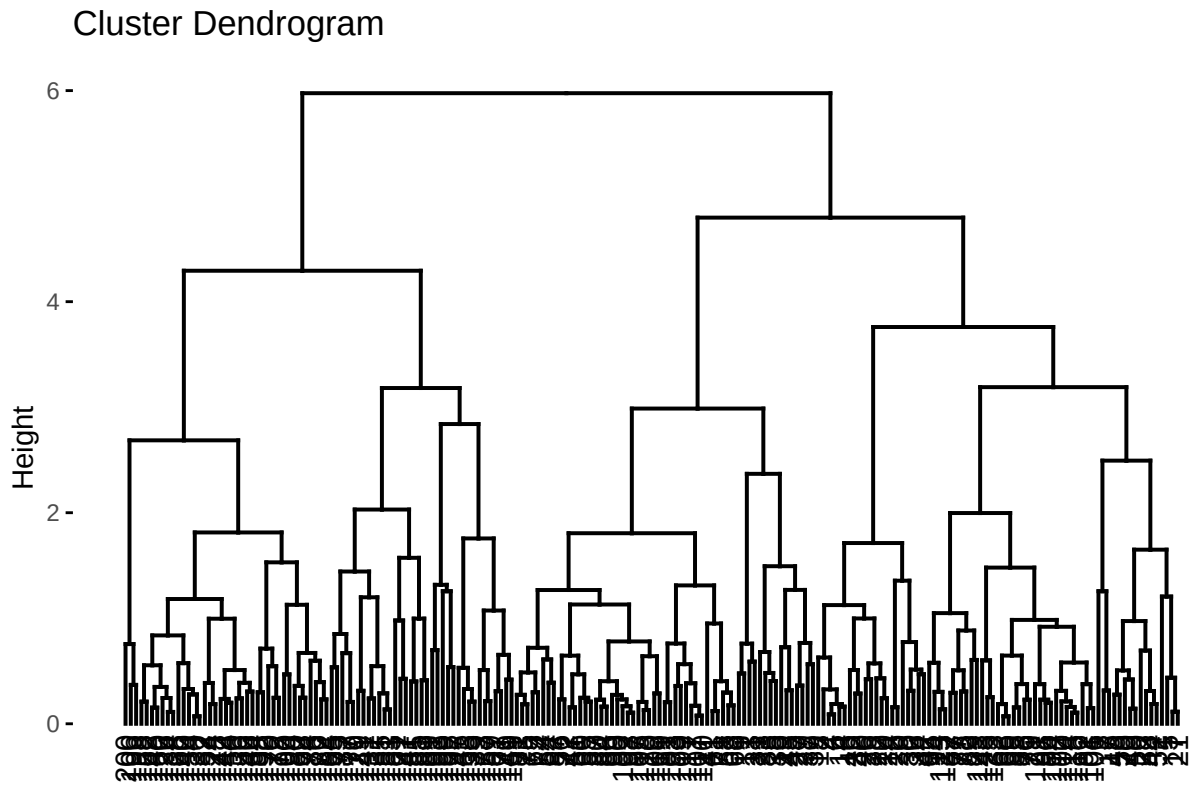


Berdasarkan koefisien silhouette, metode complete dan average memilih 5 gerombol, sedangkan metode centroid dan ward masing-masing memilih 2 dan 6 gerombol. Untuk saat ini, kita akan mencoba menggunakan 5 gerombol dengan metode complete (Jika dua metode linkage memilih banyaknya gerombol yang sama, gerombol yang terbentuk akan relatif mirip, oleh karena itu bisa pilih salah satu).

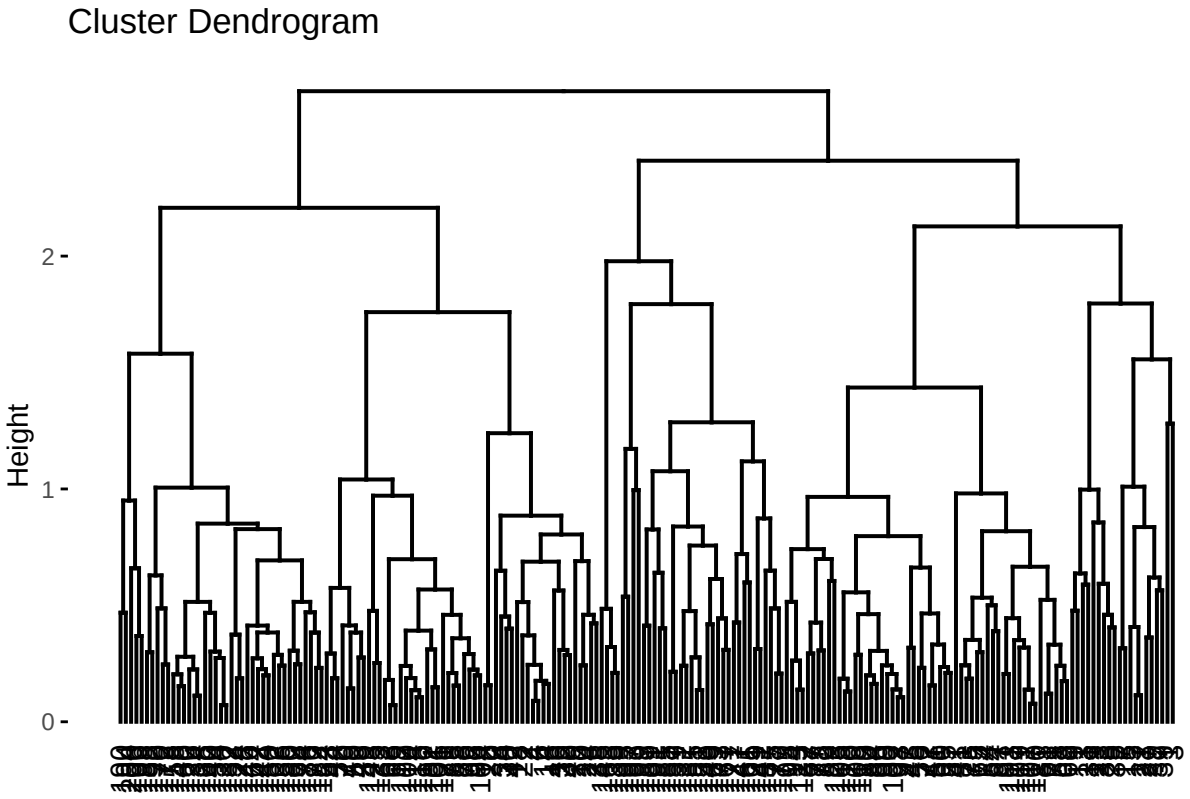
Menggunakan dendogram

Penggunaan dendogram untuk data yang memiliki amatan yang banyak mungkin tidak efektif karena memilih gerombol dengan dendogram dilakukan secara visual.

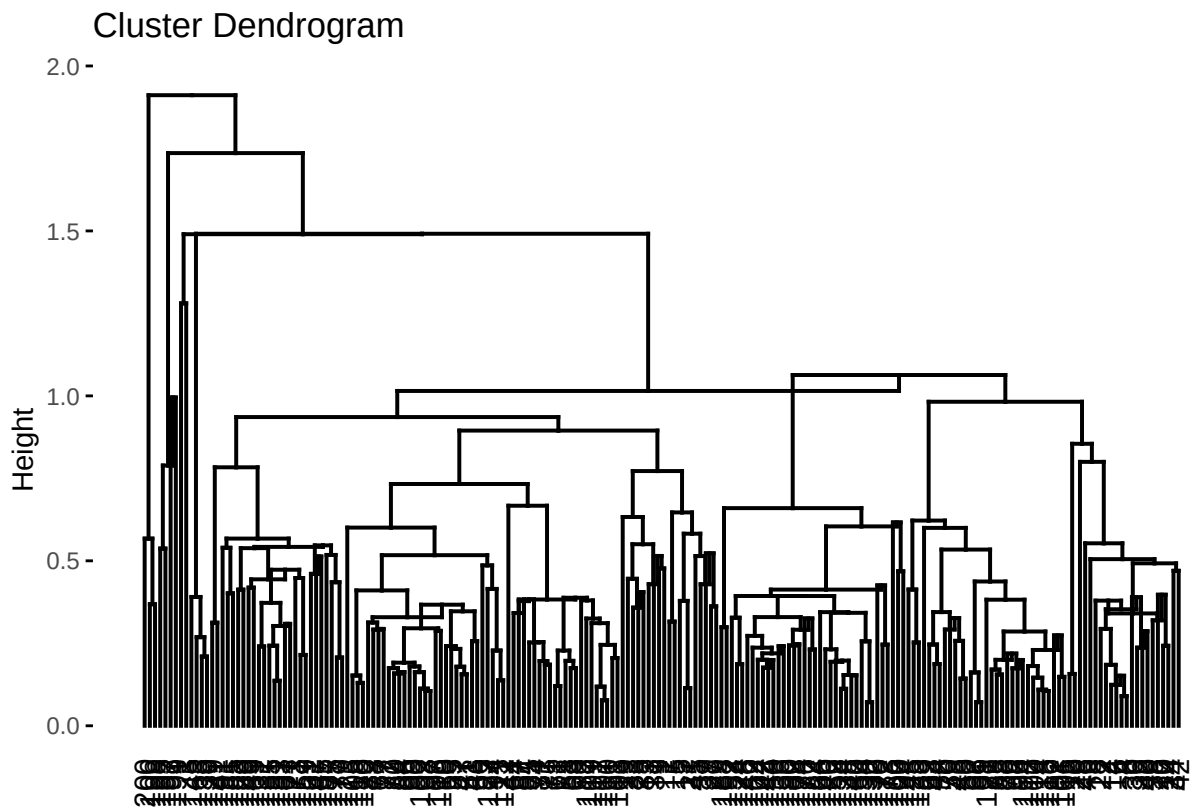
```
linkage_methods <- c("complete","average","centroid","ward.D")
hc_mall_dend <- lapply(linkage_methods, function(i)
  hclust(dist(data_mall_standardize,method = 'euclidean'),method = i)
  )
#complete
fviz_dend(hc_mall_dend[[1]])
```



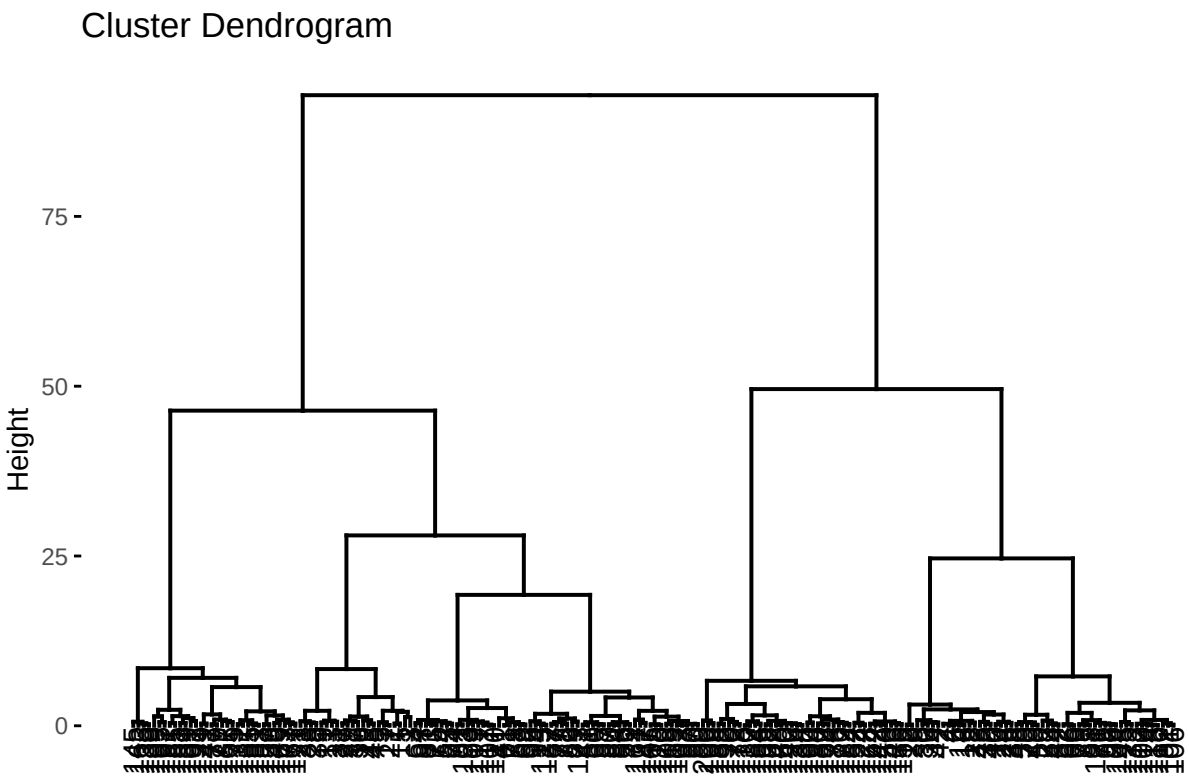

```
#average
fviz_dend(hc_mall_dend[[2]])
```




```
#centroid
fviz_dend(hc_mall_dend[[3]])
```




```
#ward
fviz_dend(hc_mall_dend[[4]])
```



Jika diperhatikan dari keempat dendogram pada masing-masing metode linkage, banyaknya gerombol yang terbentuk sama seperti menggunakan keofisien silhouette diatas.

3. Menerapkan Hierarchical Clustering

```
hc_mall <- eclust(data_mall,stand = TRUE,FUNcluster = "hclust",k=5,hc_method = "complete",hc_metric = "
hc_mall$cluster
```

```
##   [1] 1 2 1 2 1 2 1 2 3 2 3 2 3 2 1 2 1 2 3 2 1 2 3 2 3 2 3 1 3 2 3 2 3 2 3
##  [36] 2 3 2 3 2 3 2 3 1 3 2 3 1 1 1 3 1 1 3 3 3 3 3 1 3 3 1 3 3 3 1 1 3 1 1
##  [71] 3 3 3 3 3 1 1 1 1 3 3 1 3 3 1 3 3 1 1 3 3 1 3 1 1 1 3 1 3 1 1 3 3 1 3
## [106] 1 3 3 3 3 3 1 1 1 1 1 3 3 3 3 1 1 1 4 1 4 5 4 5 4 5 4 1 4 5 4 5 4 5 4
## [141] 5 4 1 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5
## [176] 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4
```

4. Interprestasi Gerombol yang terbentuk

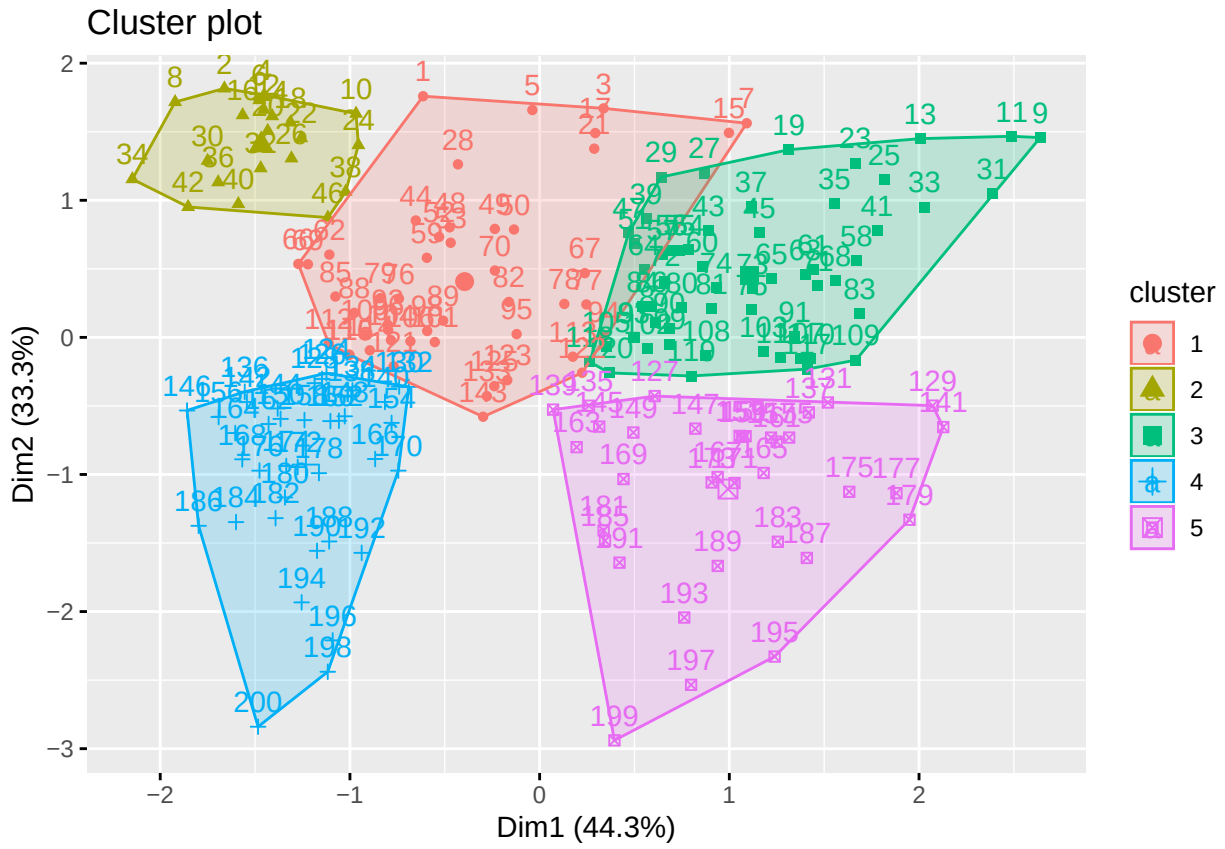
Coba lakukan interpretasi gerombol seperti metode kmeans diatas

```
aggregate(data_mall,by =list(gerombol=hc_mall$cluster),
          FUN = mean)
```

```
##   gerombol      Age Annual.Income Spending.Score
## 1        1 28.35417      50.29167       45.93750
## 2        2 24.80952      25.61905       80.23810
## 3        3 55.33333      47.31579       41.08772
## 4        4 32.69231      86.53846       82.12821
## 5        5 41.68571      88.22857       17.28571
```

Scatterplot

```
fviz_cluster(hc_mall)
```



Interpretasi dua komponen utama bisa dilihat dengan akar cirinya.

```
pca_mall <- prcomp(data_mall_standardize)
pca_mall$rotation
```

```
##                       PC1         PC2         PC3
## Age             0.70638235 -0.03014116 0.707188441
## Annual.Income  -0.04802398 -0.99883160 0.005397916
## Spending.Score -0.70619946  0.03777499 0.707004506
```