Hierarchical Clustering

Tutorial ini adalah bagian dari mata kuliah Fundamental Sains Data, Prodi Sarjana Informatika, UII.

Data USArrests Pada contoh ini kita akan menggunakan sebuah data set `USArrest`, yang disediakan oleh package `datasets`. Data ini berisi statistik pengangkapan penjahat per 100.000 penduduk pada 50 negara bagian di U.S. pada tahun 1973, yang berdasarkan tiga jenis tindak kejahatan: penyerangan (`Assault`), pembunuhan (`Murder`), dan pemerkosaan (`Rape`), serta presentase penduduk yang tinggal di daerah urban.

Pertama, mari kita lihat lebih detail tentang data tersebut.

```
USArrests
```

```
##                Murder Assault UrbanPop Rape
## Alabama          13.2     236       58 21.2
## Alaska           10.0     263       48 44.5
## Arizona           8.1     294       80 31.0
## Arkansas          8.8     190       50 19.5
## California        9.0     276       91 40.6
## Colorado          7.9     204       78 38.7
## Connecticut       3.3     110       77 11.1
## Delaware          5.9     238       72 15.8
## Florida          15.4     335       80 31.9
## Georgia          17.4     211       60 25.8
## Hawaii            5.3      46       83 20.2
## Idaho             2.6     120       54 14.2
## Illinois         10.4     249       83 24.0
## Indiana           7.2     113       65 21.0
## Iowa              2.2      56       57 11.3
## Kansas            6.0     115       66 18.0
## Kentucky          9.7     109       52 16.3
## Louisiana        15.4     249       66 22.2
## Maine             2.1      83       51  7.8
## Maryland         11.3     300       67 27.8
## Massachusetts     4.4     149       85 16.3
## Michigan         12.1     255       74 35.1
## Minnesota         2.7      72       66 14.9
## Mississippi      16.1     259       44 17.1
## Missouri          9.0     178       70 28.2
## Montana           6.0     109       53 16.4
## Nebraska          4.3     102       62 16.5
## Nevada           12.2     252       81 46.0
## New Hampshire     2.1      57       56  9.5
## New Jersey        7.4     159       89 18.8
## New Mexico       11.4     285       70 32.1
## New York         11.1     254       86 26.1
## North Carolina   13.0     337       45 16.1
```

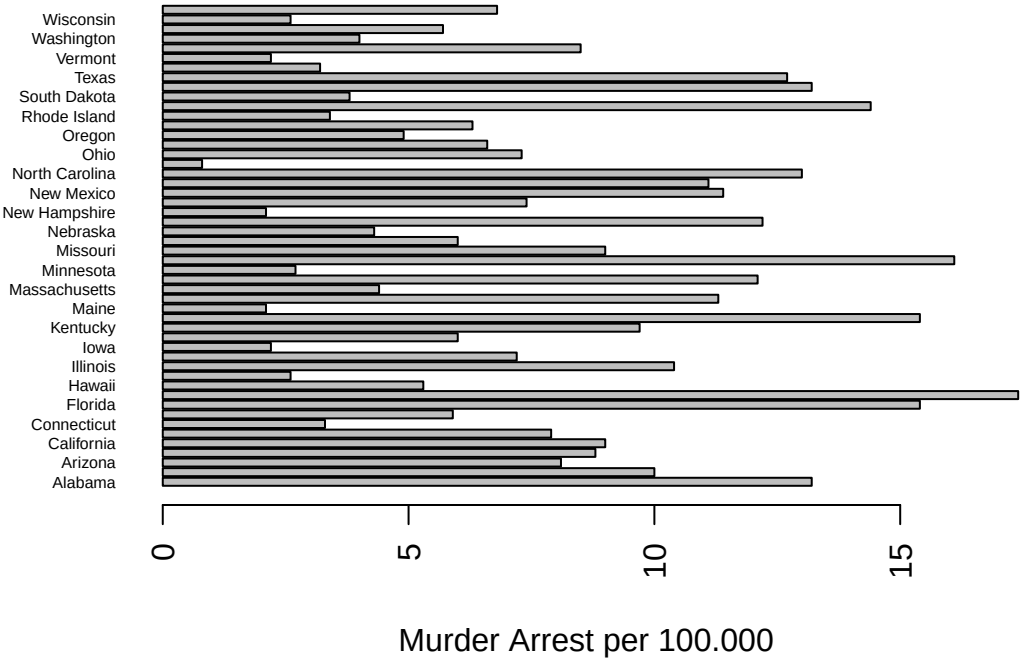
```
## North Dakota      0.8      45       44  7.3
## Ohio              7.3     120       75 21.4
## Oklahoma          6.6     151       68 20.0
## Oregon            4.9     159       67 29.3
## Pennsylvania      6.3     106       72 14.9
## Rhode Island      3.4     174       87  8.3
## South Carolina   14.4     279       48 22.5
## South Dakota      3.8      86       45 12.8
## Tennessee        13.2     188       59 26.9
## Texas            12.7     201       80 25.5
## Utah              3.2     120       80 22.9
## Vermont           2.2      48       32 11.2
## Virginia          8.5     156       63 20.7
## Washington        4.0     145       73 26.2
## West Virginia     5.7      81       39  9.3
## Wisconsin         2.6      53       66 10.8
## Wyoming           6.8     161       60 15.6
```

```
summary(USArrests)
```

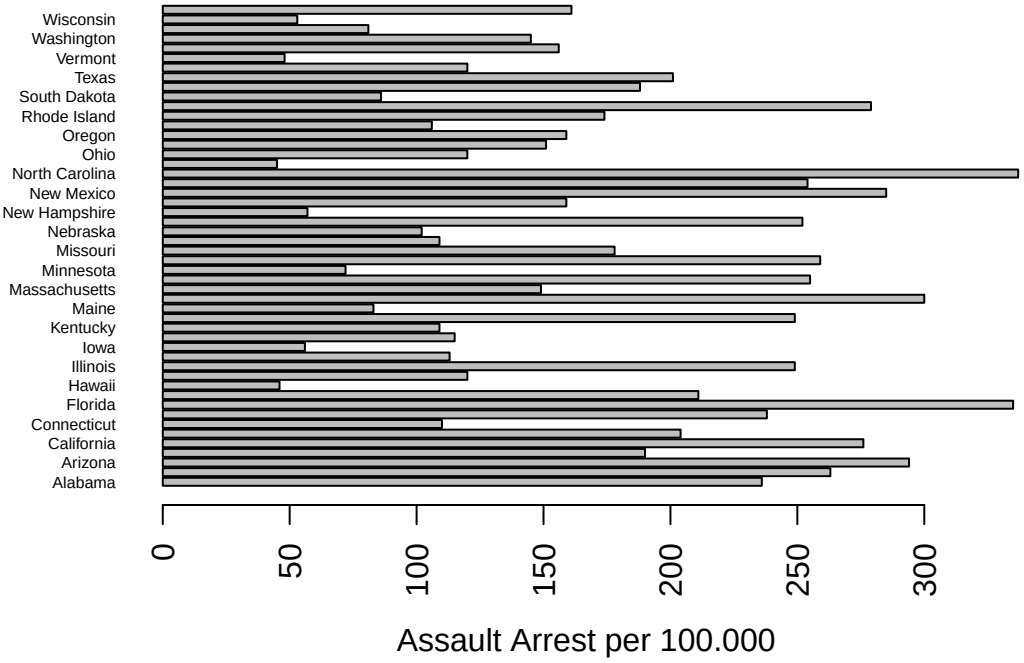
```
##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

Sebuah alternatif untuk mendapatkan visualisasi dari setiap variabel adalah dengan menggunakan `barplot()`. Baris 1 membuat sebuat data frame dari data set `USArrest` dan menambahkan satu kolom berisi nama negara bagian (state). Baris 2 untuk mengubah orientasi label pada sumbu vertikal sehingga daapt terbaca secara horizontal (lihat nama negara bagian), dan Baris 3 untuk mengatur margin, sehingga possi barplot lebih di tengah. Baris 4 membuat barplot untuk variabel `Murder`; baris-barise selanjutnya membuat `barplot` untuk variabel lain.

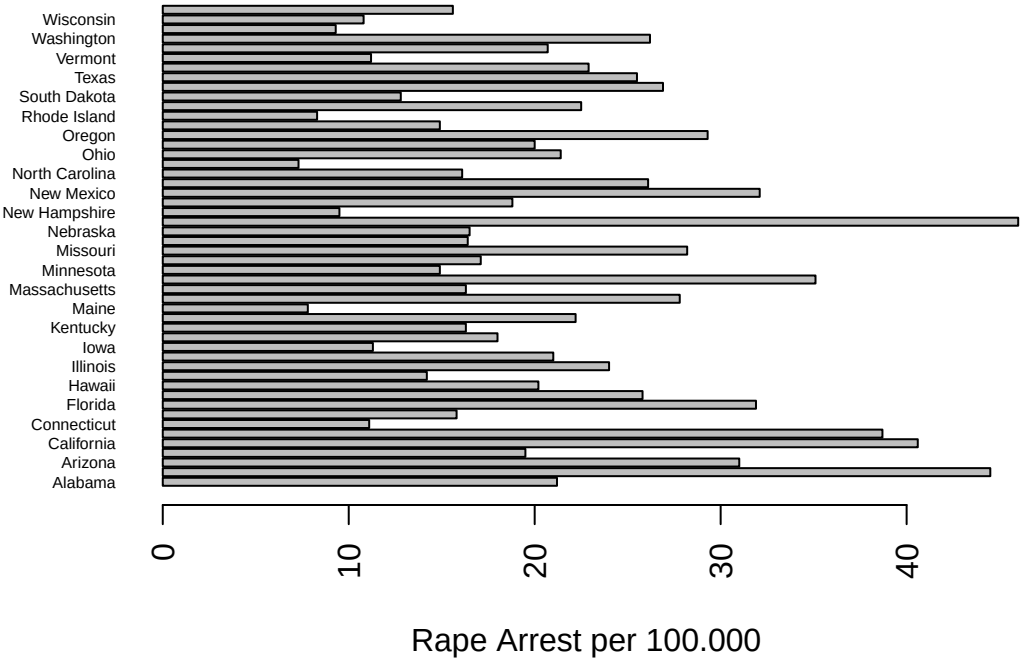
```
df <- data.frame(States=rownames(USArrests), USArrests)
par(las=2) # make label text perpendicular to axis
par(mar=c(5,8,4,2)) # increase y-axis margin.
barplot(df$Murder, names.arg = df$States, horiz = TRUE, cex.names = 0.5, xlab = "Murder Arrest per 100.
```



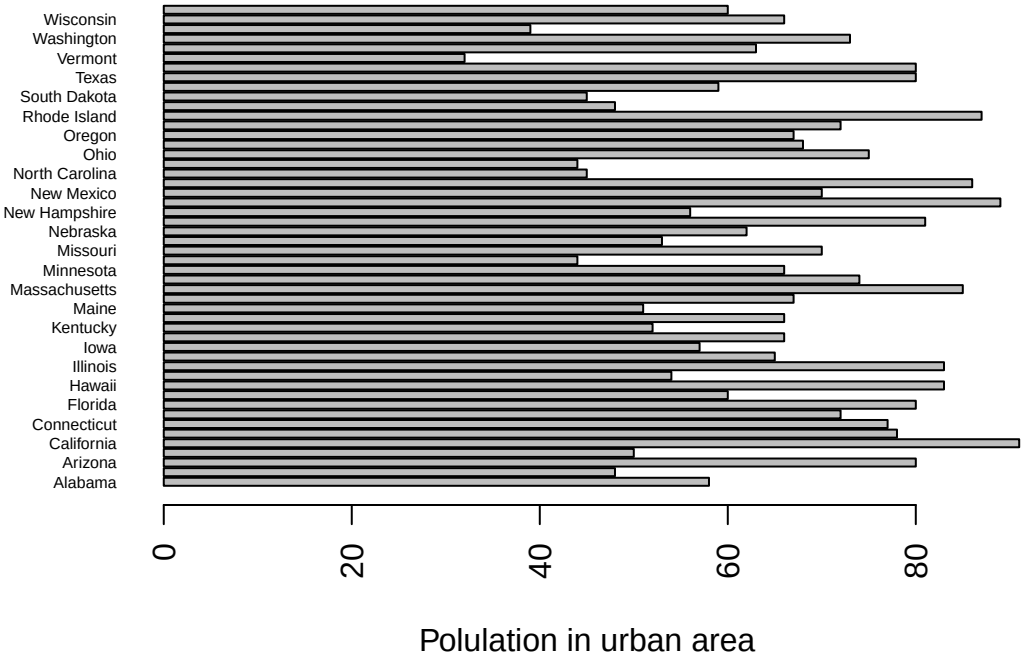

```r
barplot(df$Assault, names.arg = df$States, horiz = TRUE, cex.names = 0.5, xlab = "Assault Arrest per 10
```




```r
barplot(df$Rape, names.arg = df$States, horiz = TRUE, cex.names = 0.5, xlab = "Rape Arrest per 100.000")
```



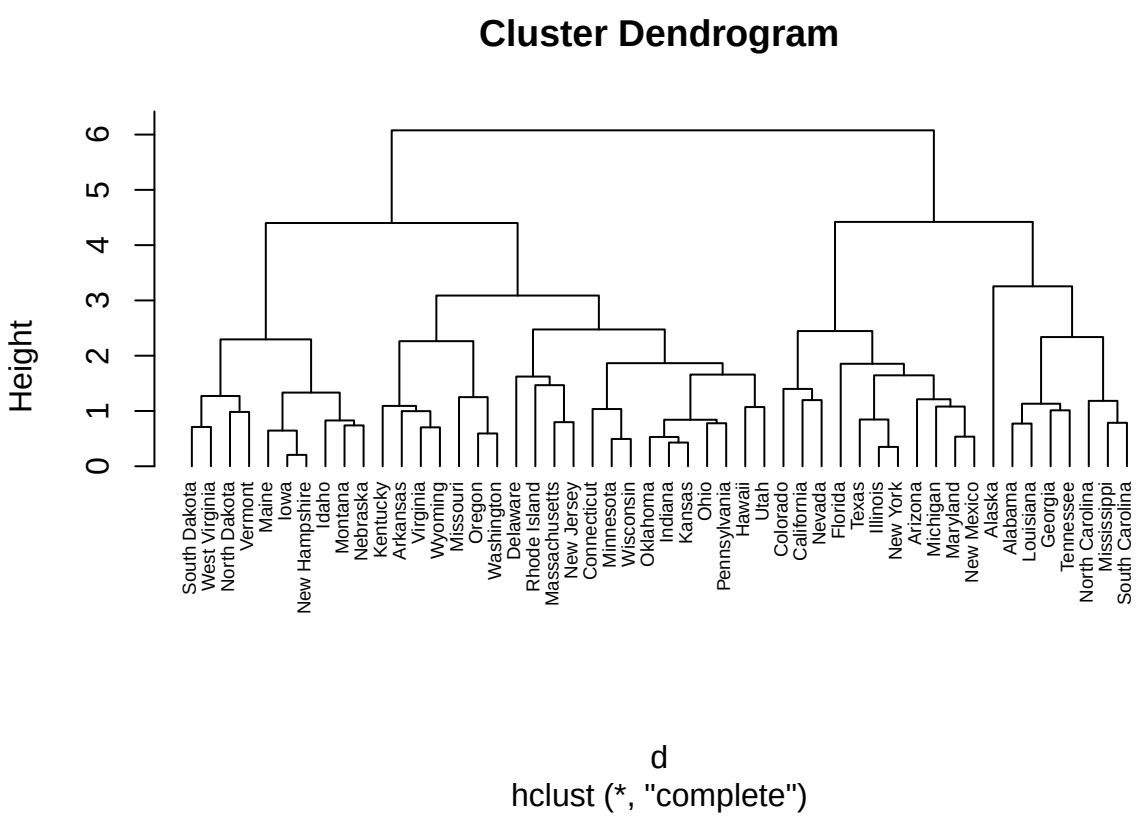

```r
barplot(df$UrbanPop, names.arg = df$States, horiz = TRUE, cex.names = 0.5, xlab = "Polulation in urban a
```



Perhatikan sebaran setiap variabel berdasarakan empat barplot di atas, untuk mendapatkan insight tentang data.

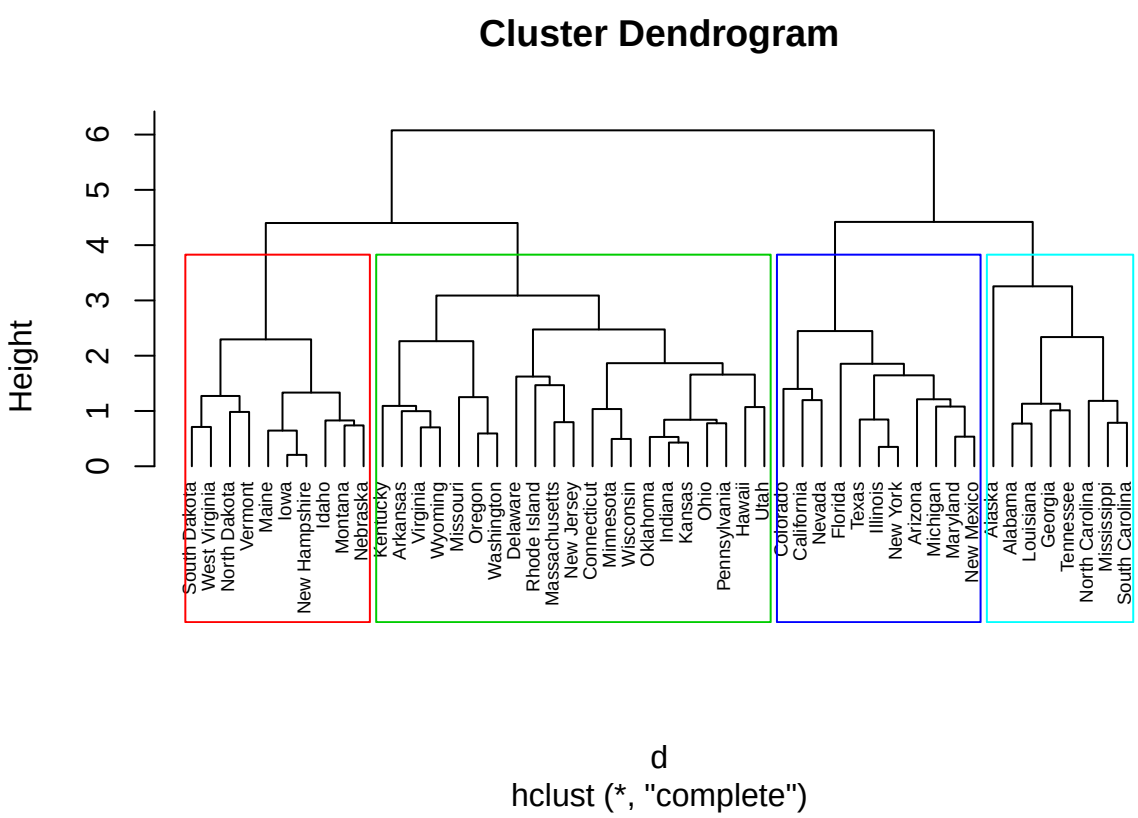
Hierarchical Clustering Di sini kita akan melakukan hierarchical clustering menggunakan fungsi `hclust`. Di sini kita hanya ingin menggunakan empat variabel (tanpa variabel `States`), maka kita hilangkan variabel tersebut lewat Baris 1. Fungsi `hclust` menerima matriks jarak (dissimilarity meassure dari setiap pasang variabel), maka kita menghitung matriks tersebut lewat Baris 2. Pada Baris 3, kita melakukan hieararchical clustering dengan metode `complete linkage` (lihat slide materi). Rekan-rekan bisa mencoba dengan `single`, `average`, dan metode lain (lihat `?hclust` untuk lebih jelasnya). Baris 4 membuat plot dendrogram dari hasil clustering; parameter `cex` emngatur besar font untuk label pada sumbu x, `hang` mengatur posisi label terhadap sumbu y.

```
1 df <- scale(df[, 2:5])
2 d <- dist(df, method = "euclidean")
3 clusters <- hclust(d, method = "complete" )
4 plot(clusters, cex = 0.6, hang = -1)
```



Kita dapat menggunakan fungsi `rect.hclust()` untuk menggambar kotak pada sejumlah klaster yang kita inginkan. Sebagai contoh, kita ingin melihat 4 klaster dari hasil clustering di atas. Maka kita set `k=4`. Parameter `border` mengatur warna kotak dari setiap klaster.

```
plot(clusters, cex = 0.6, hang = -1)
rect.hclust(clusters, k = 4, border = 2:5)
```



Latihan Cobalah hiearchical clustering pada data `USArrests` menggunakan metode `single` dan `average`. Amati perubahannya. Kemudian dengan masing-masing metode, lihat 4 (atau ubah nilai ini) klaster menggunakan `rect.hclust`, dan bandingkan juga keanggotaan setiap klaster, dari masing-masing metode.