**Praktikum 10**

**ANALISIS CLUSTER**

Dataset :

|  |  |  |
| --- | --- | --- |
| observation | Income | Education |
| s1 | 5 | 5 |
| s2 | 6 | 6 |
| s3 | 15 | 14 |
| s4 | 16 | 15 |
| s5 | 25 | 20 |
| s6 | 30 | 19 |

Script k-mean clustering

|  |
| --- |
| > data\_cluster\_steven=read.delim("clipboard")  > klaster <- kmeans(data\_cluster\_steven[, 2:3], 3, nstart = 20)  > klaster  K-means clustering with 3 clusters of sizes 2, 2, 2  Cluster means:  Income Education  1 5.5 5.5  2 15.5 14.5  3 27.5 19.5  Clustering vector:  [1] 1 1 2 2 3 3  Within cluster sum of squares by cluster:  [1] 1 1 13  (between\_SS / total\_SS = 97.9 %)  Available components:  [1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss"  [7] "size" "iter" "ifault"  > table(klaster$cluster, data\_cluster\_steven$observation)    s1 s2 s3 s4 s5 s6  1 1 1 0 0 0 0  2 0 0 1 1 0 0  3 0 0 0 0 1 1 |

Output dan interpretasi :

|  |
| --- |
|  |

Dataset iris

|  |
| --- |
| > library(ggplot2)  > ggplot(iris, aes(Petal.Length, Petal.Width, color = Species)) + geom\_point()    > set.seed(20)  > irisCluster <- kmeans(iris[, 3:4], 3, nstart = 20)  > irisCluster  K-means clustering with 3 clusters of sizes 50, 52, 48  Cluster means:  Petal.Length Petal.Width  1 1.462000 0.246000  2 4.269231 1.342308  3 5.595833 2.037500  Clustering vector:  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  [46] 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 3 2 2 2 2 2 2  [91] 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3  [136] 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3  Within cluster sum of squares by cluster:  [1] 2.02200 13.05769 16.29167  (between\_SS / total\_SS = 94.3 %)  Available components:  [1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss"  [7] "size" "iter" "ifault"  > table(irisCluster$cluster, iris$Species)  setosa versicolor virginica  1 50 0 0  2 0 48 4  3 0 2 46  > irisCluster$cluster <- as.factor(irisCluster$cluster)  > ggplot(iris, aes(Petal.Length, Petal.Width, color = irisCluster$cluster)) + geom\_point() |

Dataset USArrests

|  |
| --- |
| > data("USArrests")  > df <- scale(USArrests)  > head(df, n = 3)  Murder Assault UrbanPop Rape  Alabama 1.24256408 0.7828393 -0.5209066 -0.003416473  Alaska 0.50786248 1.1068225 -1.2117642 2.484202941  Arizona 0.07163341 1.4788032 0.9989801 1.042878388  > library(factoextra)  > fviz\_nbclust(df, kmeans, method = "wss") +  + geom\_vline(xintercept = 4, linetype = 2) |

|  |
| --- |
| > set.seed(123)  > km.res <- kmeans(df, 4, nstart = 25)  > print(km.res)  K-means clustering with 4 clusters of sizes 13, 16, 13, 8  Cluster means:  Murder Assault UrbanPop Rape  1 -0.9615407 -1.1066010 -0.9301069 -0.96676331  2 -0.4894375 -0.3826001 0.5758298 -0.26165379  3 0.6950701 1.0394414 0.7226370 1.27693964  4 1.4118898 0.8743346 -0.8145211 0.01927104  Clustering vector:  Alabama Alaska Arizona Arkansas California Colorado  4 3 3 4 3 3  Connecticut Delaware Florida Georgia Hawaii Idaho  2 2 3 4 2 1  Illinois Indiana Iowa Kansas Kentucky Louisiana  3 2 1 2 1 4  Maine Maryland Massachusetts Michigan Minnesota Mississippi  1 3 2 3 1 4  Missouri Montana Nebraska Nevada New Hampshire New Jersey  3 1 1 3 1 2  New Mexico New York North Carolina North Dakota Ohio Oklahoma  3 3 4 1 2 2  Oregon Pennsylvania Rhode Island South Carolina South Dakota Tennessee  2 2 2 4 1 4  Texas Utah Vermont Virginia Washington West Virginia  3 2 1 2 2 1  Wisconsin Wyoming  1 2  Within cluster sum of squares by cluster:  [1] 11.952463 16.212213 19.922437 8.316061  (between\_SS / total\_SS = 71.2 %)  Available components:  [1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss"  [7] "size" "iter" "ifault"  > aggregate(USArrests, by=list(cluster=km.res$cluster), mean)  cluster Murder Assault UrbanPop Rape  1 1 3.60000 78.53846 52.07692 12.17692  2 2 5.65625 138.87500 73.87500 18.78125  3 3 10.81538 257.38462 76.00000 33.19231  4 4 13.93750 243.62500 53.75000 21.41250    > dd <- cbind(USArrests, cluster = km.res$cluster)  > head(dd)  Murder Assault UrbanPop Rape cluster  Alabama 13.2 236 58 21.2 4  Alaska 10.0 263 48 44.5 3  Arizona 8.1 294 80 31.0 3  Arkansas 8.8 190 50 19.5 4  California 9.0 276 91 40.6 3  Colorado 7.9 204 78 38.7 3  > km.res$cluster  Alabama Alaska Arizona Arkansas California Colorado  4 3 3 4 3 3  Connecticut Delaware Florida Georgia Hawaii Idaho  2 2 3 4 2 1  Illinois Indiana Iowa Kansas Kentucky Louisiana  3 2 1 2 1 4  Maine Maryland Massachusetts Michigan Minnesota Mississippi  1 3 2 3 1 4  Missouri Montana Nebraska Nevada New Hampshire New Jersey  3 1 1 3 1 2  New Mexico New York North Carolina North Dakota Ohio Oklahoma  3 3 4 1 2 2  Oregon Pennsylvania Rhode Island South Carolina South Dakota Tennessee  2 2 2 4 1 4  Texas Utah Vermont Virginia Washington West Virginia  3 2 1 2 2 1  Wisconsin Wyoming  1 2  > km.res$size  [1] 13 16 13 8  > km.res$centers  Murder Assault UrbanPop Rape  1 -0.9615407 -1.1066010 -0.9301069 -0.96676331  2 -0.4894375 -0.3826001 0.5758298 -0.26165379  3 0.6950701 1.0394414 0.7226370 1.27693964  4 1.4118898 0.8743346 -0.8145211 0.01927104    > fviz\_cluster(km.res, data = df,  + palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),  + ellipse.type = "euclid", # Concentration ellipse  + star.plot = TRUE, # Add segments from centroids to items  + repel = TRUE, # Avoid label overplotting (slow)  + ggtheme = theme\_minimal()  + ) |

Output dan interpretasi :

|  |
| --- |
|  |

Dataset kasus mandiri (praktikan mencari kasus sendiri) :

|  |
| --- |
| > data\_pangan\_steven=read.delim("clipboard")  > klaster <- kmeans(data\_pangan\_steven[, 2:7], 7, nstart = 20)  > klaster\_pangan <- kmeans(data\_pangan\_steven[, 2:7], 7, nstart = 20)  > klaster\_pangan  K-means clustering with 7 clusters of sizes 3, 2, 8, 2, 5, 9, 3  Cluster means:  Beras Daging.ayam Daging.sapi Telur.ayam Cabai.merah Gula.pasir  1 11483.33 45616.67 104133.3 24416.67 29033.33 13500.00  2 12775.00 42400.00 112125.0 28900.00 46075.00 15425.00  3 10662.50 28818.75 109693.8 21412.50 22518.75 13306.25  4 12300.00 42100.00 124675.0 28675.00 25875.00 14800.00  5 12180.00 30930.00 125350.0 21120.00 42210.00 13470.00  6 11550.00 32616.67 119016.7 20861.11 31772.22 12705.56  7 11750.00 30700.00 112233.3 24233.33 44466.67 13266.67  Clustering vector:  [1] 6 5 3 5 3 4 5 5 5 7 6 6 6 6 3 1 1 4 6 1 3 7 2 2 3 6 6 7 6 3 3 3  Within cluster sum of squares by cluster:  [1] 217898333 28170000 206556250 172138750 262709000 251756667 51166667  (between\_SS / total\_SS = 80.5 %)  Available components:  [1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss"  [7] "size" "iter" "ifault"  >table(klaster\_pangan$cluster, data\_pangan\_steven$Provinsi)  Aceh Bali Banten Bengkulu DI Yogyakarta DKI Jakarta Gorontalo Jambi Jawa Barat Jawa Tengah  1 0 0 0 0 0 0 0 0 0 0  2 0 0 0 0 0 0 0 0 0 0  3 0 1 0 0 0 0 1 0 0 0  4 0 0 0 0 0 0 0 0 0 0  5 1 0 0 0 0 1 0 0 0 0  6 0 0 1 1 1 0 0 1 1 1  7 0 0 0 0 0 0 0 0 0 0    Jawa Timur Kalimantan Barat Kalimantan Selatan Kalimantan Timur Kalimantan Utara  1 0 0 0 0 1  2 0 0 0 0 0  3 2 0 0 0 0  4 0 0 1 0 0  5 0 1 0 0 0  6 0 0 0 1 0  7 0 0 0 0 0    Kepulauan Bangka Belitung Lampung Maluku Nusa Tenggara Barat Nusa Tenggara Timur Papua  1 0 0 0 1 1 0  2 0 0 1 0 0 0  3 0 0 0 0 0 0  4 0 0 0 0 0 1  5 0 0 0 0 0 0  6 0 1 0 0 0 0  7 1 0 0 0 0 0    Papua Barat Riau Sulawesi Barat Sulawesi Selatan Sulawesi Tengah Sulawesi Tenggara  1 0 0 0 0 0 0  2 1 0 0 0 0 0  3 0 0 1 1 1 1  4 0 0 0 0 0 0  5 0 0 0 0 0 0  6 0 0 0 0 0 0  7 0 1 0 0 0 0    Sulawesi Utara Sumatera Barat Sumatera Selatan Sumatera Utara  1 0 0 0 0  2 0 0 0 0  3 0 0 0 0  4 0 0 0 0  5 0 1 0 1  6 0 0 1 0  7 1 0 0 0  ggplot(data\_pangan\_steven, aes(Beras, Daging.ayam,Daging.sapi,Telur.ayam,Cabai.merah,Gula.pasir, color = klaster\_pangan$cluster)) + geom\_point()    > df\_pangan <- scale(data\_pangan\_steven[, 2:7])  > head(df\_pangan, n = 3)  Beras Daging.ayam Daging.sapi Telur.ayam Cabai.merah Gula.pasir  [1,] -0.2536536 -0.007304901 -0.1502604 -0.5860745 -0.3392329 -0.82461934  [2,] 1.5340006 0.039446464 1.3752751 -0.2540107 1.2632123 0.04055505  [3,] -0.3019686 0.366706020 -0.9548677 -0.6582623 -1.1348915 -0.34877342  >fviz\_nbclust(df\_pangan, kmeans, method = "wss") + geom\_vline(xintercept = 4, linetype = 2)    > km.res\_pangan <- kmeans(df\_pangan, 4, nstart = 25)  > print(km.res\_pangan)  K-means clustering with 4 clusters of sizes 12, 10, 4, 6  Cluster means:  Beras Daging.ayam Daging.sapi Telur.ayam Cabai.merah Gula.pasir  1 -0.02013124 -0.4228726 0.5277553 -0.3971832 0.5004744 -0.6335600  2 -0.94938937 -0.4755977 -0.7018992 -0.3940550 -1.0403027 -0.1541092  3 1.38905569 0.4465729 1.2497563 -0.1926511 0.1044997 0.2460340  4 0.69654097 1.3406928 -0.7188496 1.5795587 0.6632227 1.3599460  Clustering vector:  [1] 1 3 2 1 2 4 1 1 3 1 1 1 2 1 2 2 4 3 3 4 2 4 4 4 2 1 1 1 1 2 2 2  Within cluster sum of squares by cluster:  [1] 27.12564 19.79724 7.75453 26.59988  (between\_SS / total\_SS = 56.3 %)  Available components:  [1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss"  [7] "size" "iter" "ifault"  > dd <- cbind(data\_pangan\_steven, cluster = km.res\_pangan$cluster)  > head(dd)  Provinsi Beras Daging.ayam Daging.sapi Telur.ayam Cabai.merah Gula.pasir cluster  1 Lampung 11300 33600 114400 20650 29500 12450 1  2 DKI Jakarta 13150 33900 126250 21800 43900 13450 3  3 Bali 11250 36000 108150 20400 22350 13000 2  4 Kalimantan Barat 13100 27400 123150 21650 47800 11750 1  5 Sulawesi Selatan 9950 24450 106000 20100 20550 12950 2  6 Papua 12200 42500 121200 34750 31750 16350  >aggregate(data\_pangan\_steven[,2:7], by=list(cluster=km.res\_pangan$cluster), mean)  cluster Beras Daging.ayam Daging.sapi Telur.ayam Cabai.merah Gula.pasir  1 1 11181.82 34018.18 115227.3 21654.55 32722.73 12900.00  2 2 12866.67 31508.33 121916.7 21516.67 39616.67 13000.00  3 3 10561.11 27827.78 113666.7 21327.78 26172.22 13455.56  4 4 12458.33 43833.33 112691.7 27750.00 34725.00 14650.00  > km.res\_pangan$size  [1] 12 10 4 6  > km.res\_pangan$centers  Beras Daging.ayam Daging.sapi Telur.ayam Cabai.merah Gula.pasir  1 -0.02013124 -0.4228726 0.5277553 -0.3971832 0.5004744 -0.6335600  2 -0.94938937 -0.4755977 -0.7018992 -0.3940550 -1.0403027 -0.1541092  3 1.38905569 0.4465729 1.2497563 -0.1926511 0.1044997 0.2460340  4 0.69654097 1.3406928 -0.7188496 1.5795587 0.6632227 1.3599460  > fviz\_cluster(km.res\_pangan, data = df\_pangan, palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"), ellipse.type = "euclid", # Concentration ellipse  + star.plot = TRUE, # Add segments from centroids to items  + repel = TRUE, # Avoid label overplotting (slow)  + ggtheme = theme\_minimal()) |

Output dan interpretasi :

|  |
| --- |
| Data tiap provinsi terbagi menjadi 4 kategori cluster utama sesuai grafik warna diatas. |

Sumber :

<https://datascienceplus.com/k-means-clustering-in-r/>

<https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/>