**Praktikum 8 – Data Analitik**

**ANALISIS KOMPONEN UTAMA (PRINCIPAL COMPONENT ANALYSIS (PCA))**

**DAN ANALISIS KLASTER**

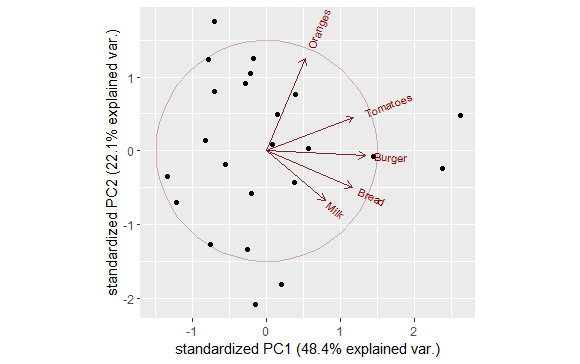
Analisis komponen utama (principal component analysis) dapat digunakan untuk membuat visualisasi atau peta posisi dari berbagai obyek dalam dimensi yang lebih rendah. Hal ini dapat terjadi karena beberapa variabel asal dari matriks data saling berkorelasi atau terjadinya multikolinieritas. Sedangkan analisis klaster digunakan (cluster analysis) dapat digunakan untuk mengelompokkan obyek berdasarkan nilai kemiripannya (similarity) berdasarkan konsep jarak antara dua titik.

Dalam tugas praktikum kali ini anda diminta melakukan analisis data terhadap data harga komoditas di berbagai kota di Amerika Serikat. Berikan interpretasi terhadap setiap output yang dihasilkan dan gunakan inisial nama anda dalam file data yang diberikan.

# PCA

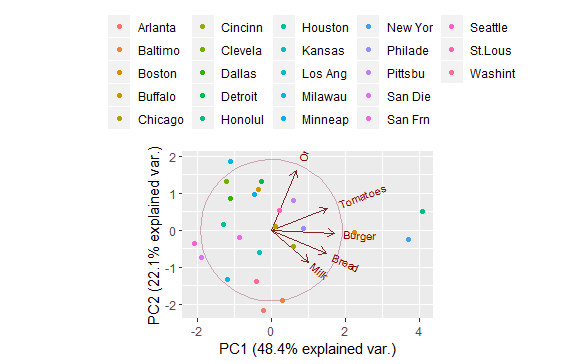
|  |
| --- |
| > library(RMySQL)  Warning messages:  1: package ‘RMySQL’ was built under R version 3.4.4  2: package ‘DBI’ was built under R version 3.4.4  3: package ‘arules’ was built under R version 3.4.4  4: package ‘Matrix’ was built under R version 3.4.4  > con = dbConnect(MySQL(), user = 'root', password = '', dbname =  + 'db\_da', host = 'localhost')  > myQuery <- "select \* from komoditas;"  > data\_steven <- dbGetQuery(con, myQuery)  Warning messages:  1: In .local(conn, statement, ...) :  Decimal MySQL column 1 imported as numeric  2: In .local(conn, statement, ...) :  Decimal MySQL column 2 imported as numeric  3: In .local(conn, statement, ...) :  Decimal MySQL column 3 imported as numeric  4: In .local(conn, statement, ...) :  Decimal MySQL column 4 imported as numeric  5: In .local(conn, statement, ...) :  Decimal MySQL column 5 imported as numeric  > View(data\_steven) |

|  |  |  |  |
| --- | --- | --- | --- |
| > str(data\_steven)  'data.frame': 23 obs. of 6 variables:  $ ty : chr "Arlanta" "Baltimo" "Boston" "Buffalo" ...  $ Bread : num 24.5 26.6 29.7 22.8 26.7 25.3 22.8 23.3 24.1 29.3 ...  $ Burger : num 94.5 91 100.8 86.6 86.7 ...  $ Milk : num 73.9 67.5 61.4 65.3 62.7 63.3 52.4 62.5 51.5 80.2 ...  $ Oranges : num 80.1 74.6 104 118.4 105.9 ...  $ Tomatoes: num 41.6 53.3 59.6 51.2 51.2 45.6 46.8 41.8 52.4 61.7 ...  > data\_steven.pca=data\_steven[,2:6]  > komoditas\_stev.pca <- prcomp(data\_steven.pca, center = TRUE, scale. = TRUE)  > print(komoditas\_stev.pca)  Standard deviations (1, .., p=5):  [1] 1.5562491 1.0519527 0.8593360 0.7009800 0.4915821  Rotation (n x k) = (5 x 5):  PC1 PC2 PC3 PC4 PC5  Bread 0.4960921 -0.31116324 0.38400354 0.50926221 -0.500263314  Burger 0.5756736 -0.04237556 0.26199524 -0.02488239 0.773009430  Milk 0.3400964 -0.42779987 -0.83615624 0.04580486 0.008144644  Oranges 0.2238587 0.79723159 -0.29066455 0.47931306 -0.009064873  Tomatoes 0.5066698 0.28773392 0.01601972 -0.71287711 -0.389928880  > plot(komoditas\_stev.pca, type = "l")    > summary(komoditas\_stev.pca)  Importance of components:  PC1 PC2 PC3 PC4 PC5  Standard deviation 1.5562 1.0520 0.8593 0.70098 0.49158  Proportion of Variance 0.4844 0.2213 0.1477 0.09827 0.04833  Cumulative Proportion 0.4844 0.7057 0.8534 0.95167 1.00000   |  | | --- | | > predict(komoditas\_stev.pca, newdata=tail(data\_steven))  PC1 PC2 PC3 PC4 PC5  18 0.6037072 0.8071284 0.2265320 0.5828393 0.14339325  19 0.2257479 0.5205205 0.1305183 0.8809165 -0.06240393  20 -1.8911706 -0.7300545 0.5653872 0.9435969 -0.18954093  21 -0.8622959 -0.1935930 0.4784665 0.8343109 -0.31822364  22 -2.0904434 -0.3677658 -0.6486172 -0.5605815 -0.68562580  23 -0.3945226 -1.3977846 -0.1146839 -0.6838917 0.56682383 | |  | |  |   > hasil\_stev=predict(komoditas\_stev.pca, newdata=tail(data\_steven))  > View(hasil)    > komoditas\_stev.kota <- data\_steven[, 1]  > library(devtools)  > library(ggbiplot)  > g\_stev <- ggbiplot(komoditas\_stev.pca,ellipse = TRUE, circle = TRUE)  > g\_stev <- g\_stev + scale\_color\_discrete(name = '')  > g\_stev <- g\_stev + theme(legend.direction = 'horizontal', legend.position = 'top')  > print(g\_stev) |



Menampilkan nama kota pada gambar di atas

|  |
| --- |
| > g\_stev <- ggbiplot(komoditas\_stev.pca, obs.scale = 1, var.scale = 1, groups= komoditas\_stev.kota, circle = TRUE)  > g\_stev <- g\_stev + scale\_color\_discrete(name = '')  > g\_stev <- g\_stev + theme(legend.direction = 'horizontal', legend.position = 'top')  > print(g\_stev) |



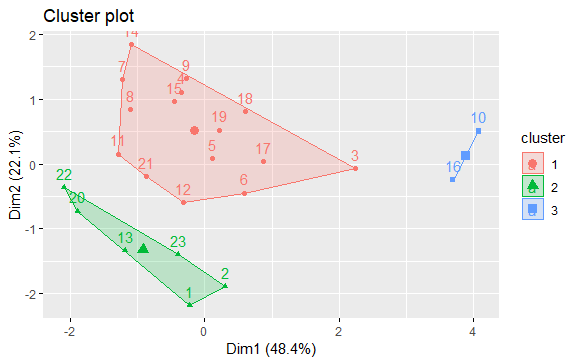
Sumber : Sharma, S. (1996) Applied Multivariate Technique.

<https://tgmstat.wordpress.com/2013/11/28/computing-and-visualizing-pca-in-r/>

# K-mean Cluster

**NOTES: *ganti nama\_sendiri atau namasendiri dengan nama kalian masing-masing***

|  |
| --- |
| > library(RMySQL)  > con = dbConnect(MySQL(), user = 'root', password = '', dbname =  + 'db\_da', host = 'localhost')  > myQuery <- "select \* from komoditas;"  > data\_steven <- dbGetQuery(con, myQuery)  Warning messages:  1: In .local(conn, statement, ...) :  Decimal MySQL column 1 imported as numeric  2: In .local(conn, statement, ...) :  Decimal MySQL column 2 imported as numeric  3: In .local(conn, statement, ...) :  Decimal MySQL column 3 imported as numeric  4: In .local(conn, statement, ...) :  Decimal MySQL column 4 imported as numeric  5: In .local(conn, statement, ...) :  Decimal MySQL column 5 imported as numeric  > View(data\_steven) |
| > library(factoextra)  > library(NbClust)  > str(data\_steven)  'data.frame': 23 obs. of 6 variables:  $ ty : chr "Arlanta" "Baltimo" "Boston" "Buffalo" ...  $ Bread : num 24.5 26.6 29.7 22.8 26.7 25.3 22.8 23.3 24.1 29.3 ...  $ Burger : num 94.5 91 100.8 86.6 86.7 ...  $ Milk : num 73.9 67.5 61.4 65.3 62.7 63.3 52.4 62.5 51.5 80.2 ...  $ Oranges : num 80.1 74.6 104 118.4 105.9 ...  $ Tomatoes: num 41.6 53.3 59.6 51.2 51.2 45.6 46.8 41.8 52.4 61.7 ...  > data\_steven.pca=data\_steven[,2:6]  > View(data\_steven.pca)    > library(NbClust)  > nb <- NbClust(data\_steven.pca, distance = "euclidean", min.nc =2, max.nc = 10, method = "complete", index ="all")  \*\*\* : The Hubert index is a graphical method of determining the number of clusters.  In the plot of Hubert index, we seek a significant knee that corresponds to a  significant increase of the value of the measure i.e the significant peak in Hubert  index second differences plot.    \*\*\* : The D index is a graphical method of determining the number of clusters.  In the plot of D index, we seek a significant knee (the significant peak in Dindex  second differences plot) that corresponds to a significant increase of the value of  the measure.    \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  \* Among all indices:  \* 8 proposed 2 as the best number of clusters  \* 9 proposed 3 as the best number of clusters  \* 1 proposed 4 as the best number of clusters  \* 1 proposed 6 as the best number of clusters  \* 2 proposed 8 as the best number of clusters  \* 1 proposed 9 as the best number of clusters  \* 2 proposed 10 as the best number of clusters  \*\*\*\*\* Conclusion \*\*\*\*\*    \* According to the majority rule, the best number of clusters is 3      \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  Warning message:  In pf(beale, pp, df2) : NaNs produced    > km.res=kmeans(data\_steven.pca,3,nstart = 25)  > fviz\_cluster(km.res, data = data\_steven.pca, geom = "point",stand = FALSE, frame.type = "norm")  Too few points to calculate an ellipse  Warning messages:  1: argument frame is deprecated; please use ellipse instead.  2: argument frame.type is deprecated; please use ellipse.type instead.    > fviz\_cluster(km.res, data = data\_steven.pca) |



Sumber : <http://www.sthda.com/english/wiki/print.php?id=239#nbclust-r-function>