PCA and K-Means Iris Datasheet

Mochammad Arie Nugroho

12/23/2021

Import Library

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(dplyr)
library(cluster)
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.1.2
```

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

Instruksi

Dataset: Iris Task:

Lakukan PCA sehingga mereduksi data Iris menjadi 2 kolom saja kemudian, lakukan K-Means Clustering dari data Iris yang telah berhasil direduksi. Setelah itu, jawablah 3 soal di bawah ini:

Soal:

a. Berapa nilai eigenvalue 1 dan eigenvalue 2 dari dataset Iris?
b. Berapa informasi yang masih bisa 'dijelaskan' oleh data yang telah direduksi?
c. Cari nilai 'k' yang optimal berdasarkan Elbow dan Silhouette method. Apakah nilai k = 3 masih merupakan nilai 'k' yang terbaik?

Kumpulkan dalam R Markdown saja (.RMD) dan jangan format lain.

Main Coding:

```
# Import data iris
data(iris)

# Hilangkan kolom 'label' bunga sehingga data 'terkesan' seperti unlabeled
df <- cbind(iris$Sepal.Length, iris$Sepal.Width, iris$Petal.Length, iris$Petal.Width)
```

```
# Buatlah PCA dari df
pca_iris = prcomp(df, center = TRUE, scale = TRUE)
```

```
# Lihat summary model PCA untuk menentukan eigenvalue 1 dan 2
summary(pca_iris)
```

```
## Importance of components:
##                          PC1    PC2     PC3     PC4
## Standard deviation     1.7084 0.9560 0.38309 0.14393
## Proportion of Variance 0.7296 0.2285 0.03669 0.00518
## Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

```
# Lakukan kuadrat pada bagian 'standard deviation' untuk melihat eigenvalue
pca_iris$sdev^2
```

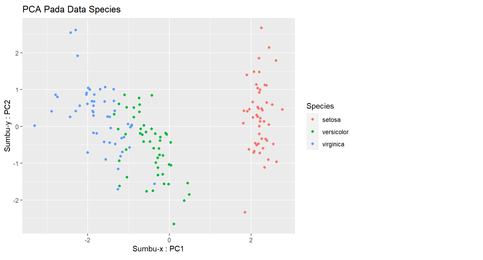
```
## [1] 2.91849782 0.91403047 0.14675688 0.02071484
```

```
# Reduksilah data iris menjadi 2 kolom saja
iris_transform = as.data.frame(-pca_iris$x[,1:2])
```

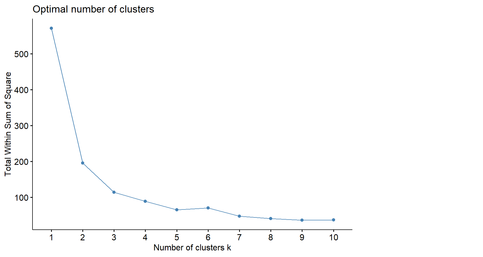
```
### Membuat dataframe baru dimana kolom pertama berisi 'row names' dari transformed data (yaitu berisi nama nama spesies)
new_data_iris = cbind(iris$Species, iris_transform)
```

```
### Membuat nama spesies sebagai sebuah kolom sendiri, dan diberi nama kolom 'Species'
rownames(new_data_iris) = NULL
colnames(new_data_iris) = c('Species', 'PC1', 'PC2')
```

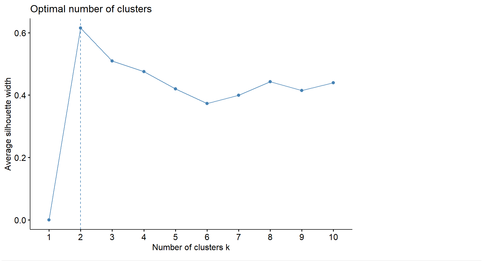
```
ggplot(new_data_iris, aes(PC1, PC2,colour=Species)) +
  geom_point() +
  xlab("Sumbu-x : PC1") +
  ylab("Sumbu-y : PC2") +
  ggtitle("PCA Pada Data Species")
```




```
# lakukan evaluasi k-means terhadap iris_transform
fviz_nbclust(iris_transform, kmeans, method = 'wss')
```




```
fviz_nbclust(iris_transform, kmeans, method = 'silhouette')
```




```
# Visualisasikan k-means clustering pada data Iris yang telah tereduksi
kmeans_iris = kmeans(iris_transform, centers = 2, nstart = 50)
fviz_cluster(kmeans_iris, data = iris_transform)
```

Summary :

a.Nilai eigenvalue 1 = 2.918497, Nilai eigenvalue 2 = 0.914030

b.informasi yang masih bisa 'dijelaskan' oleh data yang telah direduksi?

- Dengan menggunakan data yang telah direduksi dapat menampilkan 95.81% informasi dari data hanya dengan 2 principal component.
- Dari data 4 kolom, direduksi jadi 2 kolom saja. Ukuran berkurang 50%, tapi informasi yang terkandung hanya berkurang 4.19%.
- Dari visualisasi data yg didapat setelah data direduksi ditemukan bahwa species virginica dan versicolor memiliki ukuran sepal dan petal yang relatif mirip.
- Spesies setosa memiliki ukuran sepal dan petal yang jauh berbeda dengan Spesies Versicolor dan Virginica.

c.Nilai k yang optimal adalah k=2.

- berdasarkan visualisasi elbow methode nilai k yang optimal adalah 5 karena WSS mengalami peningkatan.
- berdasarkan visualisasi silhoutte methode nilai k yang optimal adalah 2 karena telah ditunjukkan oleh garis putus2 di tabel.
- Bila membandingkan metode elbow dan metode silhouette maka nilai k yang optimal adalah k=2, karena pada metode silhoutte pada k=2 memiliki nilai rata2 yang paling tinggi, dan pada metode elbow apabila nilai k=2 dapat dibilang cukup baik karena setelah nilai k=2 penambahan nilai k tidak berkurang secara signifikan. Apakah nilai k = 3 masih merupakan nilai 'k' yang terbaik? nilai k=3 pada metode silhoutte memiliki nila rata2 cukup tinggi namun masih belum bisa dikatakan yang terbaik karena tidak setinggi nilai rata2 k=2. sedangkan bila melihat visualisasi elbow nilai k=3 memang membentuk siku namun nilai k=5 masih lebih baik karena terdapat peningkatan WSS disana.