Salespeople-data

Kenneth Manuel (160419041), Jehuda Rivaldo (160419133)

3/27/2021

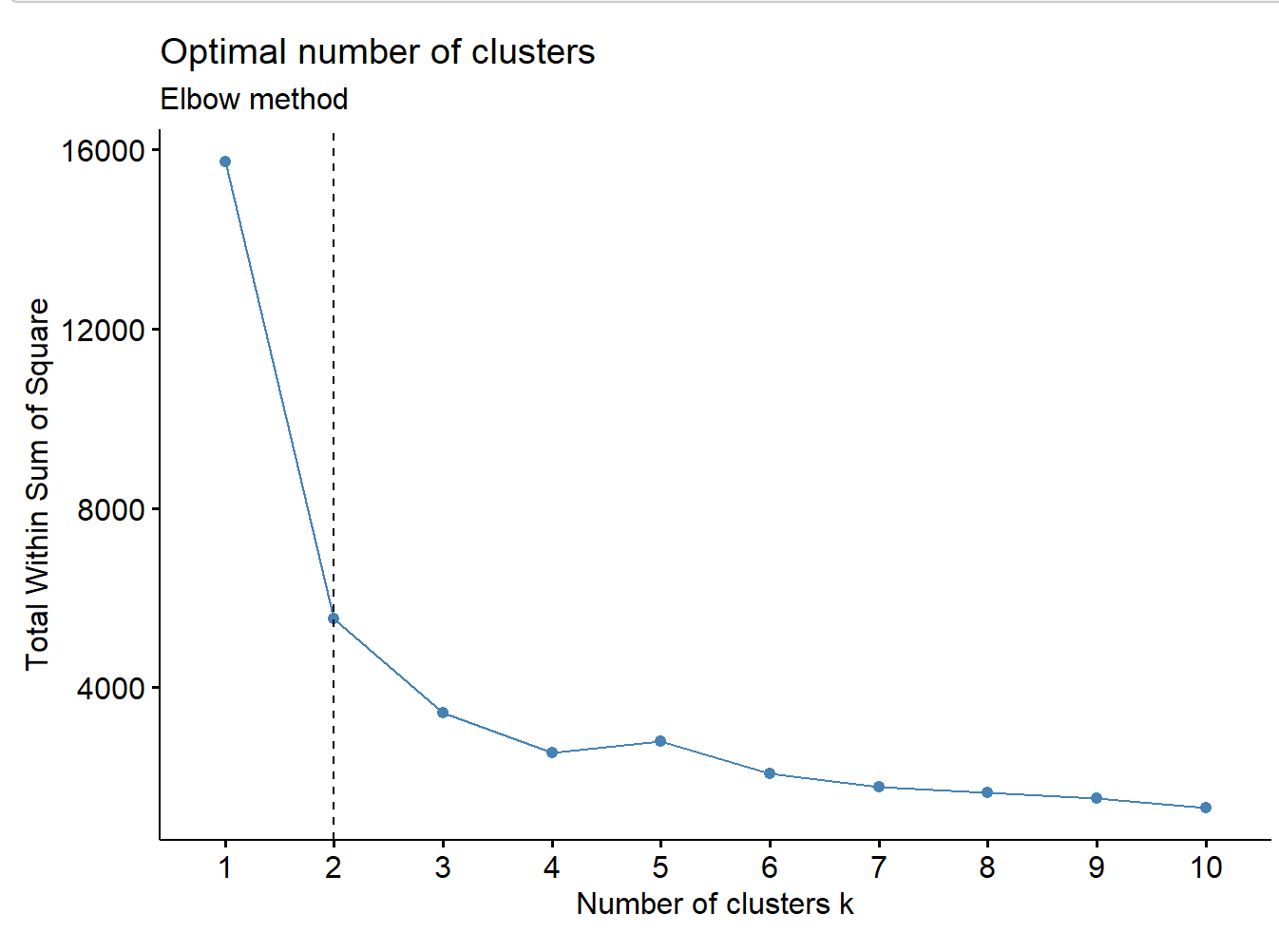
```
data <- read_excel('Salespeople-data.xlsx')
head(data)
```

Salegrow <dbl>	saleproft <dbl>	Newsale <dbl>	createst <dbl>	Mechtest <dbl>	absttest <dbl>	mathtest <dbl>
93.0	96.0	97.8	9	12	9	20
88.8	91.8	96.8	7	10	10	15
95.0	100.3	99.0	8	12	9	26
101.3	103.8	106.8	13	14	12	29
102.0	107.8	103.0	10	15	12	32
95.8	97.5	99.3	10	14	11	21

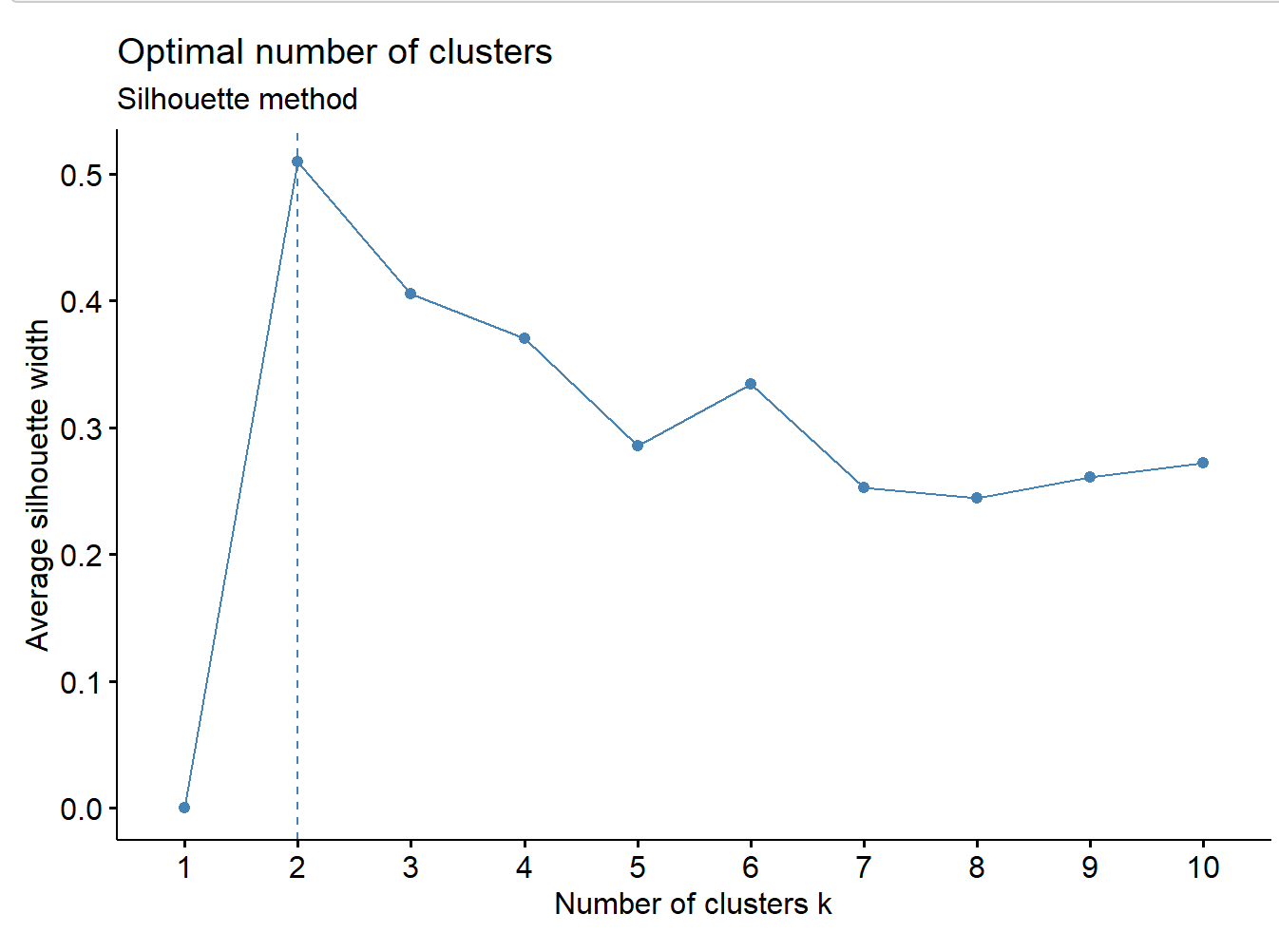
6 rows

Menentukan keputusan berapa kluster yang akan digunakan

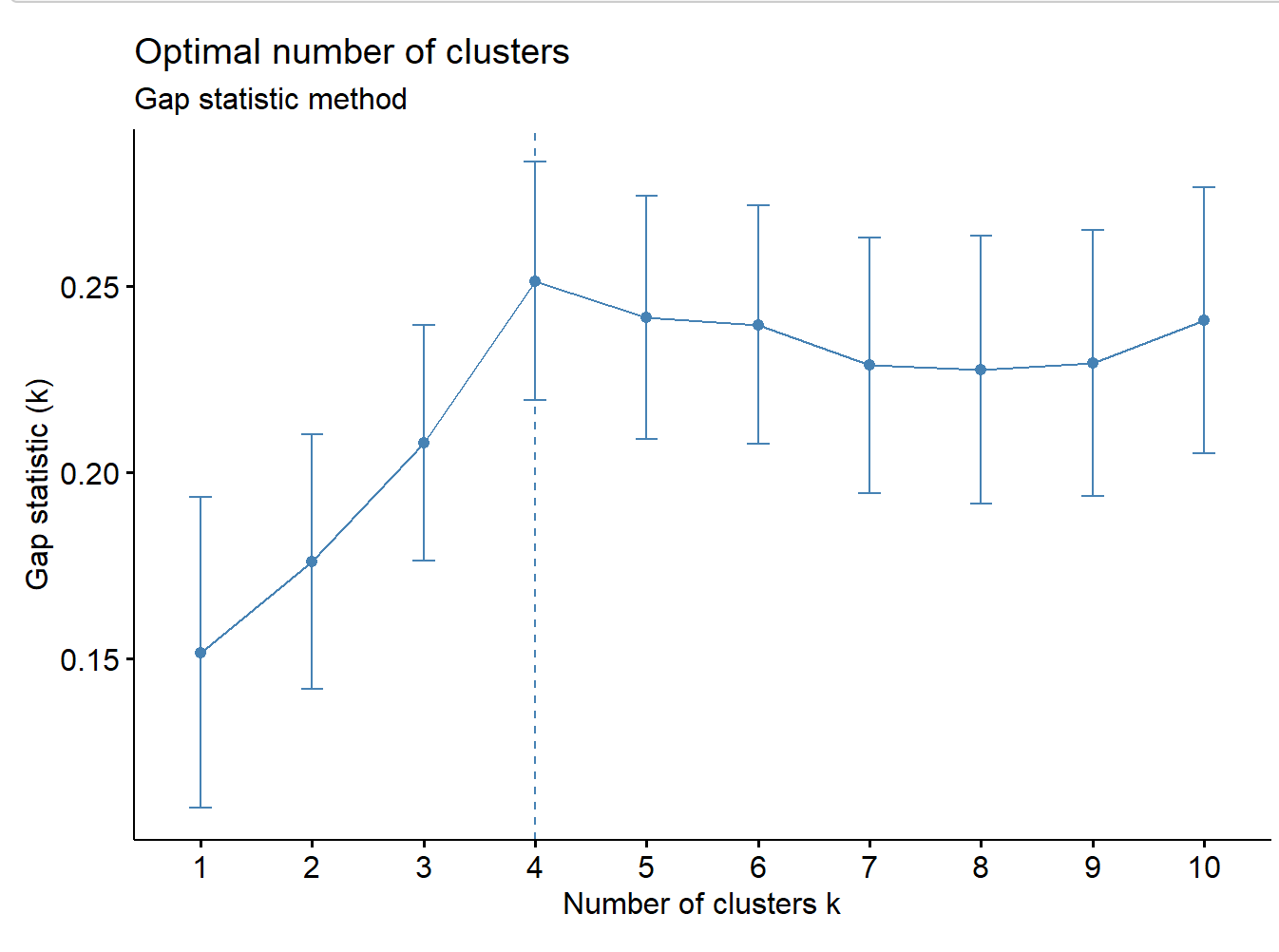
```
fviz_nbclust(data, kmeans, method = "wss") +
    geom_vline(xintercept = 2 , linetype = 2)+
  labs(subtitle = "Elbow method")
```




```
fviz_nbclust(data, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")
```




```
fviz_nbclust(data, kmeans, nstart = 25,  method = "gap_stat", nboot = 50)+
  labs(subtitle = "Gap statistic method")
```



Terlihat bahwa metode elbow dan metode silhouette menyarankan untuk menggunakan 2 cluster sedangkan gap statistics menyarankan untuk menggunakan 4 cluster.

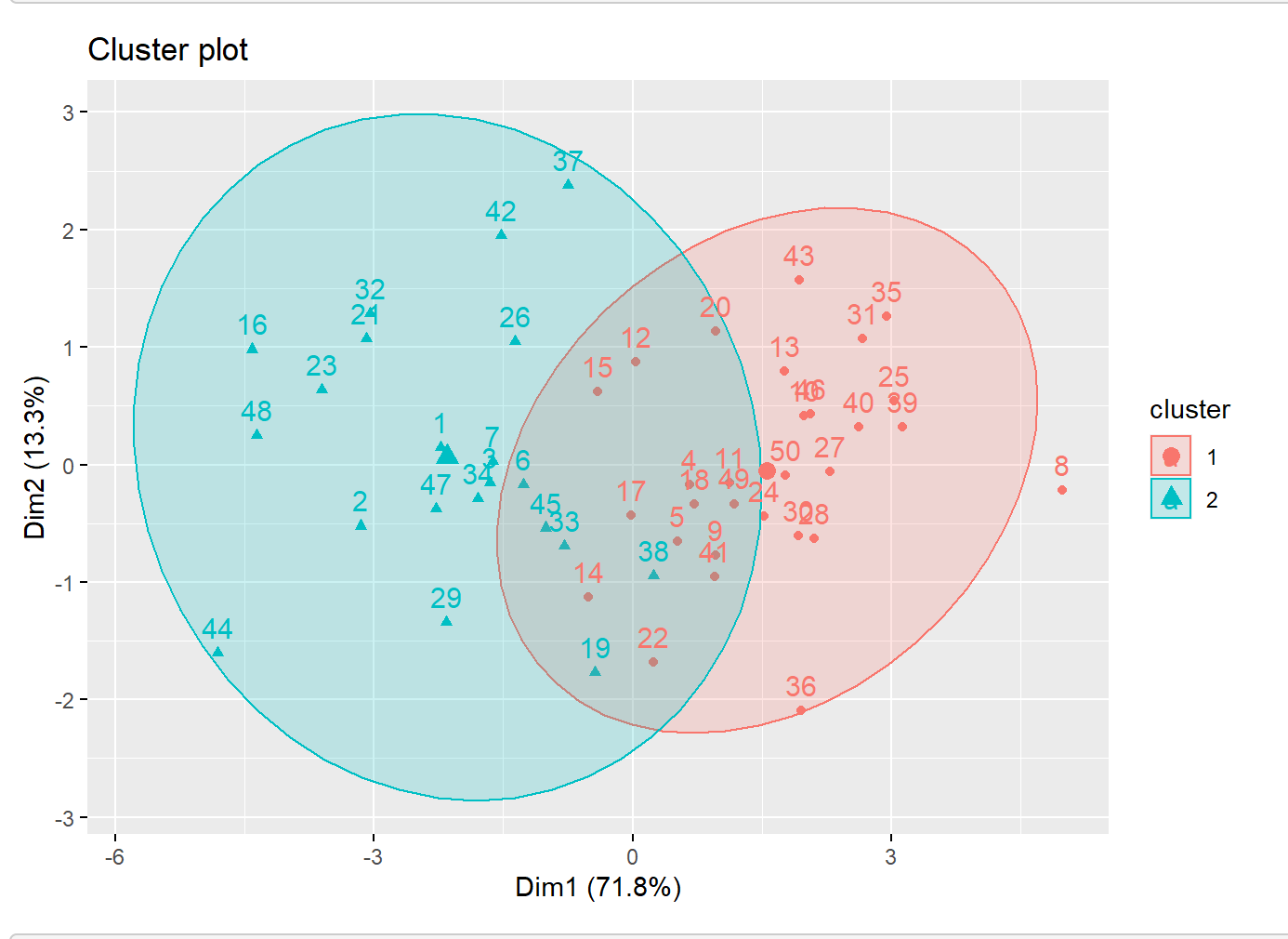
Karena dua metode yaitu silhouette dan elbow menyarankan untuk menggunakan 2 cluster maka clustering akan dibagi menjadi **2 cluster**

K-means clustering

```
km.res <- eclust(data, "kmeans", k = 2,
                 nstart = 25, graph = FALSE)
# k-means group number of each observation
km.res$cluster
```

```
##  [1] 2 2 2 1 1 2 2 1 1 1 1 1 1 1 2 1 1 2 1 2 1 2 1 1 2 1 1 2 1 1 2 2 2 1 1 2 2
## [39] 1 1 1 2 1 2 2 1 2 2 1 1
```

```
fviz_cluster(km.res,  ellipse.type = "norm")
```



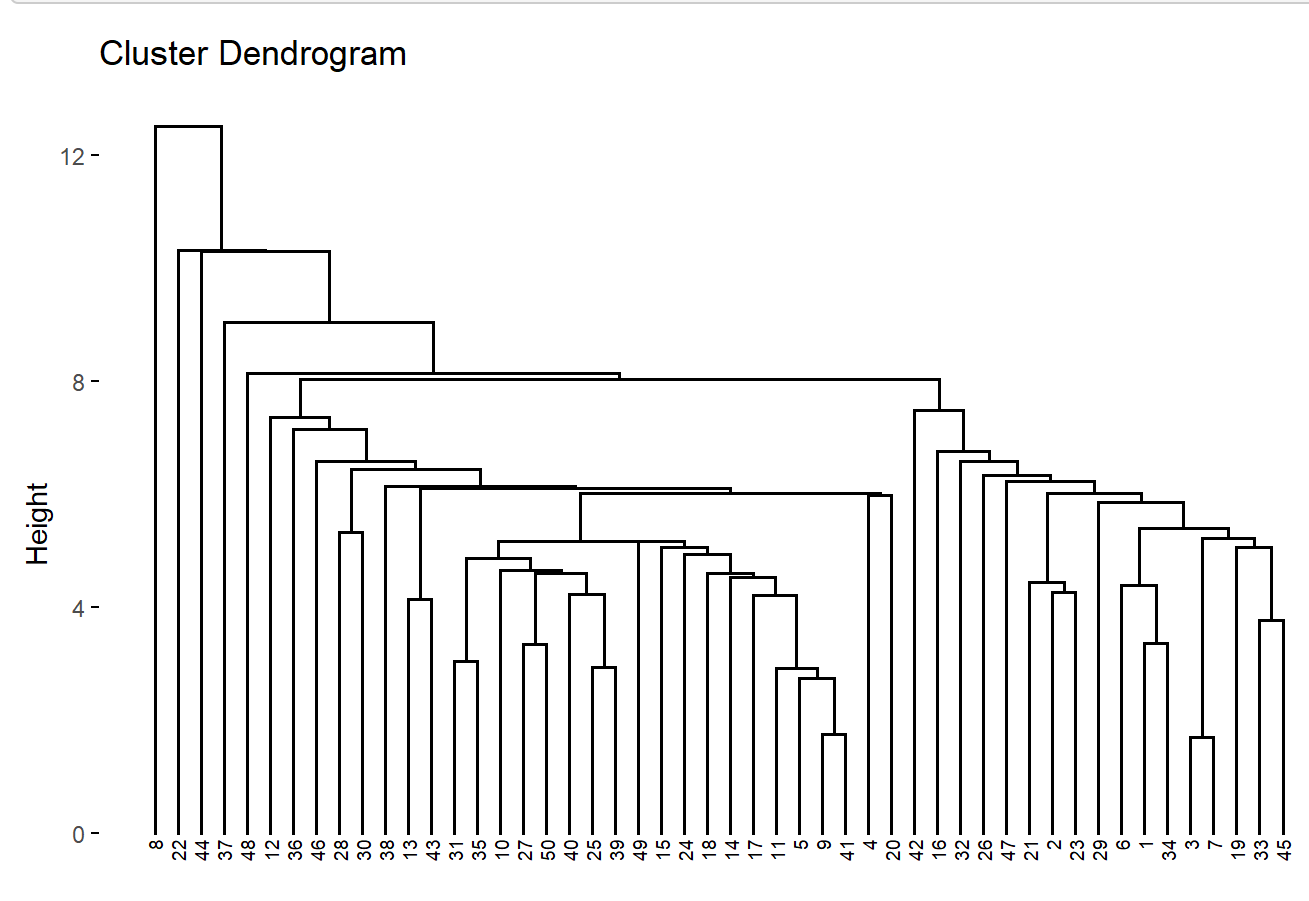

```
#fviz_silhouette(km.res)
```

Hierachical clustering dengan jarak euclidean dan metode single linkage

```
# dissimilarity matrix
res.dist <- dist(data, method = 'euclidean')
res.hc <- hclust(d = res.dist, method = 'single')
hc.cluster <- cutree(res.hc, k = 2)
hc.cluster
```

```
##  [1] 1 1 1 1 1 1 1 2 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
## [39] 1 1 1 1 1 1 1 1 1 1 1 1
```

```
fviz_dend(res.hc, rect = TRUE, show_labels = TRUE, cex = 0.5, color_labels_by_k = TRUE)
```




```
#fviz_silhouette(res.hc)
```

Comparing hierachical clustering and K-means clustering

```
km.res$cluster
```

```
##  [1] 2 2 2 1 1 2 2 1 1 1 1 1 1 1 2 1 1 2 1 2 1 2 1 1 2 1 1 2 1 1 2 2 2 1 1 2 2
## [39] 1 1 1 2 1 2 2 1 2 2 1 1
```

```
hc.cluster
```

```
##  [1] 1 1 1 1 1 1 1 2 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
## [39] 1 1 1 1 1 1 1 1 1 1 1 1
```

Attaching cluster result label to data

```
cluster.member1 <- data.frame(cluster = km.res$cluster)
data.c1 <- cbind(data, cluster.member1)

cluster.member2 <- data.frame(cluster = hc.cluster)
data.c2 <- cbind(data, cluster.member2)
```

Linear discriminant analysis hasil K-Means

Dengan mengasumsi bahwa masing - masing variabel berdistribusi normal univariate.

1. Data preparation

Split data to training and test set

```
# Split the data into training (80%) and test set (20%)
set.seed(123)
training.samples <- data.c1$cluster %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data <- data.c1[training.samples, ]
test.data <- data.c1[-training.samples, ]
```

Normalize data

```
# Estimate preprocessing parameters
preproc.param <- train.data %>%
  preProcess(method = c("center", "scale"))
# Transform the data using the estimated parameters
train.transformed <- preproc.param %>% predict(train.data)
test.transformed <- preproc.param %>% predict(test.data)
```

Compute LDA

```
model <- lda(cluster~., data = train.transformed)
model
```

```
## Call:
## lda(cluster ~ ., data = train.transformed)
##
## Prior probabilities of groups:
## -0.646418705528501   1.50831031289984
##                0.7                0.3
##
## Group means:
##                      Salegrow  saleproft    Newsale   createst   Mechtest
## -0.646418705528501  0.5430821  0.5055945  0.4661681  0.2414054  0.2591322
## 1.50831031289984   -1.2671915 -1.1797205 -1.0877255 -0.5632794 -0.6046417
##                      absttest   mathtest
## -0.646418705528501  0.3392754  0.5026441
## 1.50831031289984   -0.7916426 -1.1728362
##
## Coefficients of linear discriminants:
##                   LD1
## Salegrow  -2.08620271
## saleproft -1.33833281
## Newsale   -0.26834511
## createst  -0.06032613
## Mechtest   0.88695151
## absttest  -0.08842310
## mathtest   1.25657273
```

2. Make predictions

```
predictions <- model %>% predict(test.data)
names(predictions)
```

```
## [1] "class"     "posterior" "x"
```

3. Model accuracy

```
mean(predictions$class==test.transformed$cluster)
```

```
## [1] 0.1
```

Linear discriminant analysis hasil Hierachical clustering

Dengan mengasumsi bahwa masing - masing variabel berdistribusi normal univariate.

1. Data preparation

Split data to training and test set

```
# Split the data into training (80%) and test set (20%)
set.seed(123)
training.samples <- data.c2$cluster %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data <- data.c2[training.samples, ]
test.data <- data.c2[-training.samples, ]
```

Normalize data

```
# Estimate preprocessing parameters
preproc.param <- train.data %>%
  preProcess(method = c("center", "scale"))
# Transform the data using the estimated parameters
train.transformed <- preproc.param %>% predict(train.data)
test.transformed <- preproc.param %>% predict(test.data)
```

Compute LDA

```
model <- lda(cluster~., data = train.transformed)
model
```

```
## Call:
## lda(cluster ~ ., data = train.transformed)
##
## Prior probabilities of groups:
## -0.158113883008418   6.16644143732834
##              0.975              0.025
##
## Group means:
##                      Salegrow   saleproft     Newsale    createst    Mechtest
## -0.158113883008418 -0.0421968 -0.03730924 -0.06884791 -0.04332918 -0.04315877
## 6.16644143732834    1.6456753  1.45506023  2.68506851  1.68983814  1.68319184
##                      absttest    mathtest
## -0.158113883008418 -0.05332308 -0.04969867
## 6.16644143732834    2.07959996  1.93824821
##
## Coefficients of linear discriminants:
##                  LD1
## Salegrow  -2.3012237
## saleproft -3.1539512
## Newsale   -1.0298809
## createst   1.5560344
## Mechtest   1.3197588
## absttest   0.9053936
## mathtest   5.0205975
```

2. Make predictions

```
predictions <- model %>% predict(test.data)
names(predictions)
```

```
## [1] "class"     "posterior" "x"
```

3. Model accuracy

```
mean(predictions$class==test.transformed$cluster)
```

```
## [1] 1
```

Dilihat dari model akurasi dapat disimpulkan bahwa metode hierachical clustering dengan jarak euclidean + metode single linkage lebih akurat dibanding metode k-means.

#simpan kluster

```
clusterSales <- cbind(data, Cluster = km.res$cluster)
head(clusterSales)
```

	Salegrow <dbl>	saleproft <dbl>	Newsale <dbl>	createst <dbl>	Mechtest <dbl>	absttest <dbl>	mathtest <dbl>	Cluster <int>
1	93.0	96.0	97.8	9	12	9	20	2
2	88.8	91.8	96.8	7	10	10	15	2
3	95.0	100.3	99.0	8	12	9	26	2
4	101.3	103.8	106.8	13	14	12	29	1
5	102.0	107.8	103.0	10	15	12	32	1
6	95.8	97.5	99.3	10	14	11	21	2

6 rows

```
group <- cutree(res.hc, k=2)
HierarchySales <- cbind(data, Group = as.factor(group))
head(HierarchySales)
```

	Salegrow <dbl>	saleproft <dbl>	Newsale <dbl>	createst <dbl>	Mechtest <dbl>	absttest <dbl>	mathtest <dbl>	Group <fct>
1	93.0	96.0	97.8	9	12	9	20	1
2	88.8	91.8	96.8	7	10	10	15	1
3	95.0	100.3	99.0	8	12	9	26	1
4	101.3	103.8	106.8	13	14	12	29	1
5	102.0	107.8	103.0	10	15	12	32	1
6	95.8	97.5	99.3	10	14	11	21	1

6 rows

```
HierarchySales %>%
  group_by(Group) %>%
  summarise_all("mean")
```

Group <fct>	Salegrow <dbl>	saleproft <dbl>	Newsale <dbl>	createst <dbl>	Mechtest <dbl>	absttest <dbl>	mathtest <dbl>
1	98.71429	106.3082	102.5551	11.08163	14.06122	10.46939	29.32653
2	110.80000	122.0000	115.3000	18.00000	20.00000	15.00000	51.00000

2 rows

##iv. Buatkan analisis diskriminan untuk memperoleh: #a. Prosentase ketepatan klasifikasi 50 slaesman ke hasil iii.a. maupun ke hasil iii.b.

#b. fungsi diskriminan berdasarkan hasil iii.a. maupun ke hasil iii.b.

```
lda(Cluster~., data = clusterSales)
```

```
## Call:
## lda(Cluster ~ ., data = clusterSales)
##
## Prior probabilities of groups:
##    1    2
## 0.58 0.42
##
## Group means:
##    Salegrow saleproft   Newsale createst Mechtest  absttest mathtest
## 1 104.07931 113.80345 105.90690 12.82759 15.68966 11.413793 37.13793
## 2  91.88095  96.70476  98.53333  9.00000 12.09524  9.380952 19.57143
##
## Coefficients of linear discriminants:
##                  LD1
## Salegrow  -0.2355171
## saleproft -0.2280978
## Newsale   -0.2413533
## createst   0.1330236
## Mechtest   0.2773746
## absttest   0.1525134
## mathtest   0.1723275
```

```
lda(Group~., data = HierarchySales)
```

```
## Call:
## lda(Group ~ ., data = HierarchySales)
##
## Prior probabilities of groups:
##    1    2
## 0.98 0.02
##
## Group means:
##    Salegrow saleproft  Newsale createst Mechtest absttest mathtest
## 1  98.71429  106.3082 102.5551 11.08163 14.06122 10.46939 29.32653
## 2 110.80000  122.0000 115.3000 18.00000 20.00000 15.00000 51.00000
##
## Coefficients of linear discriminants:
##                  LD1
## Salegrow  -0.4387502
## saleproft -0.3523443
## Newsale   -0.1098142
## createst   0.3371347
## Mechtest   0.4602539
## absttest   0.3701557
## mathtest   0.5176323
```