1_clustering.R

win10

2021-05-14

```
# ############################################################################
rm(list = ls())
options(digits = 5)
# if (!is.null(dev.list())){dev.off()}
# ############################################################################

# Clustering is a form of UNsupervised learning.
# k-means clustering
library(factoextra)
```

```
## Loading required package: ggplot2
```

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

```
library(cluster)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# import (iris) dataset as data-frame
df <- as.data.frame(iris)
head(df)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
#plot(df)

#The na.omit R function removes all incomplete cases of a data object
df <- na.omit(df)

#scale/normalize/standardize
```

```r
#scale(function)
#cast as data frame
#drop 5th column because non-numeric, cannot scale non-numeric
df_scaled <- as.data.frame(scale(df[,1:4]))

set.seed(123)
#kmeans(data, centers, nstart)

#start of run
centers_value = 2
run_cluster <- kmeans(df_scaled[,1:2], centers = centers_value, nstart = 5)
a_2_12 <- run_cluster$tot.withinss

run_cluster<- kmeans(df_scaled[,1:3], centers = centers_value, nstart = 5)
a_2_13 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,1:4], centers = centers_value, nstart = 5)
a_2_14 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,2:3], centers = centers_value, nstart = 5)
a_2_23 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,2:4], centers = centers_value, nstart = 5)
a_2_24 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,3:4], centers = centers_value, nstart = 5)
a_2_34 <- run_cluster$tot.withinss

center_results <- c(a_2_34, a_2_24, a_2_23, a_2_14, a_2_13, a_2_12)

center_labels <- c("a_2_34", "a_2_24", "a_2_23", "a_2_14", "a_2_13", "a_2_12")
center_2_df <- data.frame(center_labels, center_results)
center_2_df
```

```
##   center_labels center_results
## 1        a_2_34         53.808
## 2        a_2_24        148.567
## 3        a_2_23        115.881
## 4        a_2_14        220.879
## 5        a_2_13        189.199
## 6        a_2_12        165.839
```

```r
#compare to original response
table(run_cluster$cluster, iris$Species)
```

```
##
##     setosa versicolor virginica
##   1      0         50        50
##   2     50          0         0
```

```r
#end of run

#start of run
centers_value = 3
run_cluster <- kmeans(df_scaled[,1:2], centers = centers_value, nstart = 5)
```

```
a_3_12 <- run_cluster$tot.withinss
# tot.withinss = Total within-cluster sum of squares, i.e.sum(withinss).

run_cluster<- kmeans(df_scaled[,1:3], centers = centers_value, nstart = 5)
a_3_13 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,1:4], centers = centers_value, nstart = 5)
a_3_14 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,2:3], centers = centers_value, nstart = 5)
a_3_23 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,2:4], centers = centers_value, nstart = 5)
a_3_24 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,3:4], centers = centers_value, nstart = 5)
a_3_34 <- run_cluster$tot.withinss

center_results <- c(a_3_34, a_3_24, a_3_23, a_3_14, a_3_13, a_3_12)

center_labels <- c("a_3_34", "a_3_24", "a_3_23", "a_3_14", "a_3_13", "a_3_12")

center_3_df <- data.frame(center_labels, center_results)

#compare to original response
table(run_cluster$cluster, iris$Species)
```

```
##
##     setosa versicolor virginica
##   1      0          2        46
##   2      0         48         4
##   3     50          0         0
```

```
#end of run

#start of run
centers_value = 4
run_cluster <- kmeans(df_scaled[,1:2], centers = centers_value, nstart = 5)
a_4_12 <- run_cluster$tot.withinss

run_cluster<- kmeans(df_scaled[,1:3], centers = centers_value, nstart = 5)
a_4_13 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,1:4], centers = centers_value, nstart = 5)
a_4_14 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,2:3], centers = centers_value, nstart = 5)
a_4_23 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,2:4], centers = centers_value, nstart = 5)
a_4_24 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,3:4], centers = centers_value, nstart = 5)
a_4_34 <- run_cluster$tot.withinss
```

```
center_results <- c(a_4_34, a_4_24, a_4_23, a_4_14, a_4_13, a_4_12)

center_labels <- c("a_4_34", "a_4_24", "a_4_23", "a_4_14", "a_4_13", "a_4_12")

center_4_df <- data.frame(center_labels, center_results)

#end of run

#start of run
centers_value = 5
run_cluster <- kmeans(df_scaled[,1:2], centers = centers_value, nstart = 5)
a_5_12 <- run_cluster$tot.withinss

run_cluster<- kmeans(df_scaled[,1:3], centers = centers_value, nstart = 5)
a_5_13 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,1:4], centers = centers_value, nstart = 5)
a_5_14 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,2:3], centers = centers_value, nstart = 5)
a_5_23 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,2:4], centers = centers_value, nstart = 5)
a_5_24 <- run_cluster$tot.withinss

run_cluster <- kmeans(df_scaled[,3:4], centers = centers_value, nstart = 5)
a_5_34 <- run_cluster$tot.withinss

center_results <- c(a_5_34, a_5_24, a_5_23, a_5_14, a_5_13, a_5_12)

center_labels <- c("a_5_34", "a_5_24", "a_5_23", "a_5_14", "a_5_13", "a_5_12")

center_5_df <- data.frame(center_labels, center_results)

#compare to original response
table(run_cluster$cluster, iris$Species)
```

```
##
##     setosa versicolor virginica
##   1      0         24         4
##   2      0          3        20
##   3      0          0        26
##   4     50          0         0
##   5      0         23         0
```

```
#end of run

all_results <- rbind2(center_5_df, center_4_df)
all_results <- rbind2(all_results, center_3_df)
all_results <- rbind2(all_results, center_3_df)
all_results <- rbind2(all_results, center_2_df)
all_results
```

```
##    center_labels center_results
```

```
## 1         a_5_34          9.078
## 2         a_5_24         57.490
## 3         a_5_23         42.934
## 4         a_5_14         90.228
## 5         a_5_13         76.444
## 6         a_5_12         61.530
## 7         a_4_34         12.201
## 8         a_4_24         74.289
## 9         a_4_23         56.482
## 10        a_4_14        113.332
## 11        a_4_13         94.010
## 12        a_4_12         79.388
## 13        a_3_34         17.907
## 14        a_3_24         94.683
## 15        a_3_23         74.708
## 16        a_3_14        138.888
## 17        a_3_13        118.345
## 18        a_3_12        101.931
## 19        a_3_34         17.907
## 20        a_3_24         94.683
## 21        a_3_23         74.708
## 22        a_3_14        138.888
## 23        a_3_13        118.345
## 24        a_3_12        101.931
## 25        a_2_34         53.808
## 26        a_2_24        148.567
## 27        a_2_23        115.881
## 28        a_2_14        220.879
## 29        a_2_13        189.199
## 30        a_2_12        165.839
```

```
all_results_total_withinss <- all_results[order(all_results$center_results),][1,]
all_results_total_withinss
```

```
##   center_labels center_results
## 1        a_5_34          9.078
```

```
#based on results above, select run a_5_34
a_5_34
```

```
## [1] 9.078
```

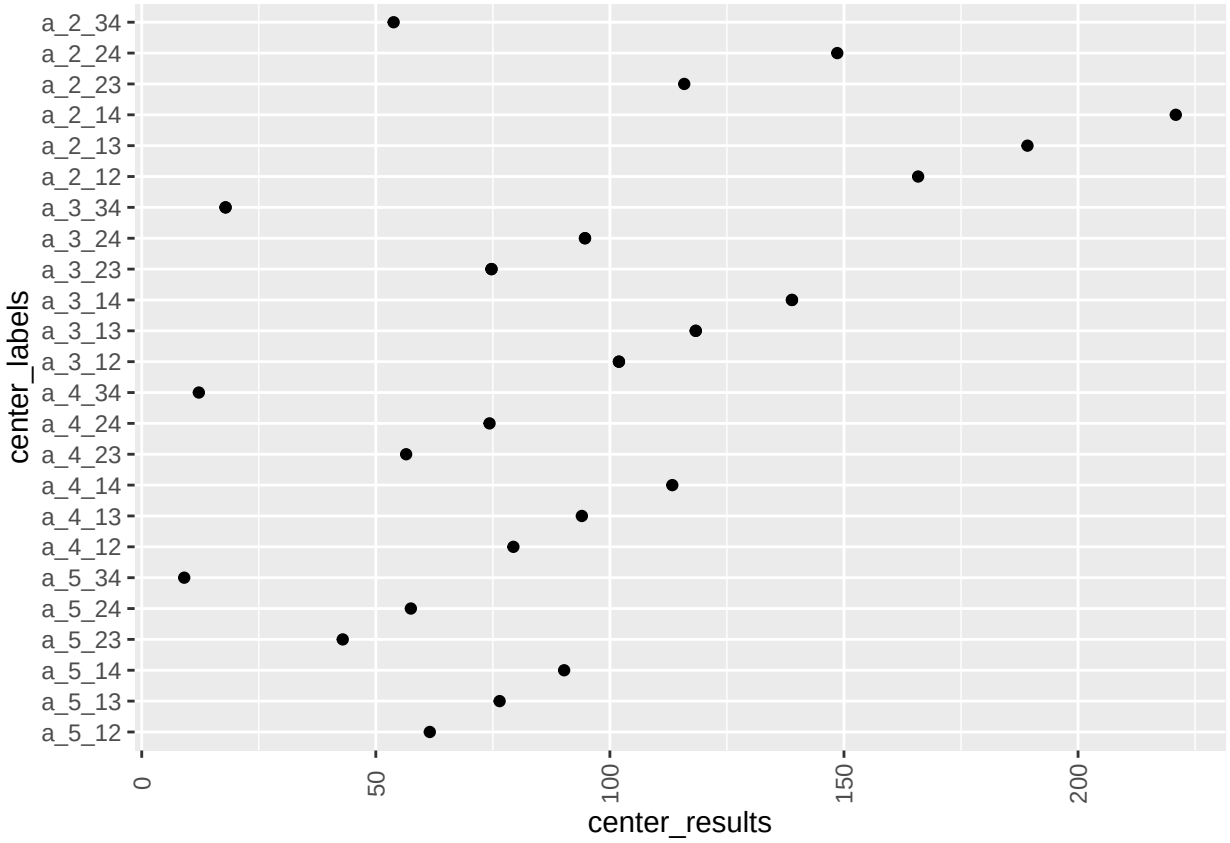
```
# rerun the clustering so that selected configuration is now active in "run_cluster" variable
run_cluster <- kmeans(df_scaled[,3:4], centers = centers_value, nstart = 5)
a_5_34 <- run_cluster$tot.withinss

#compare to original response
table(run_cluster$cluster, iris$Species)
```

```
##
##     setosa versicolor virginica
##   1      0         24         4
##   2      0          0        26
##   3      0         23         0
##   4     50          0         0
##   5      0          3        20
```

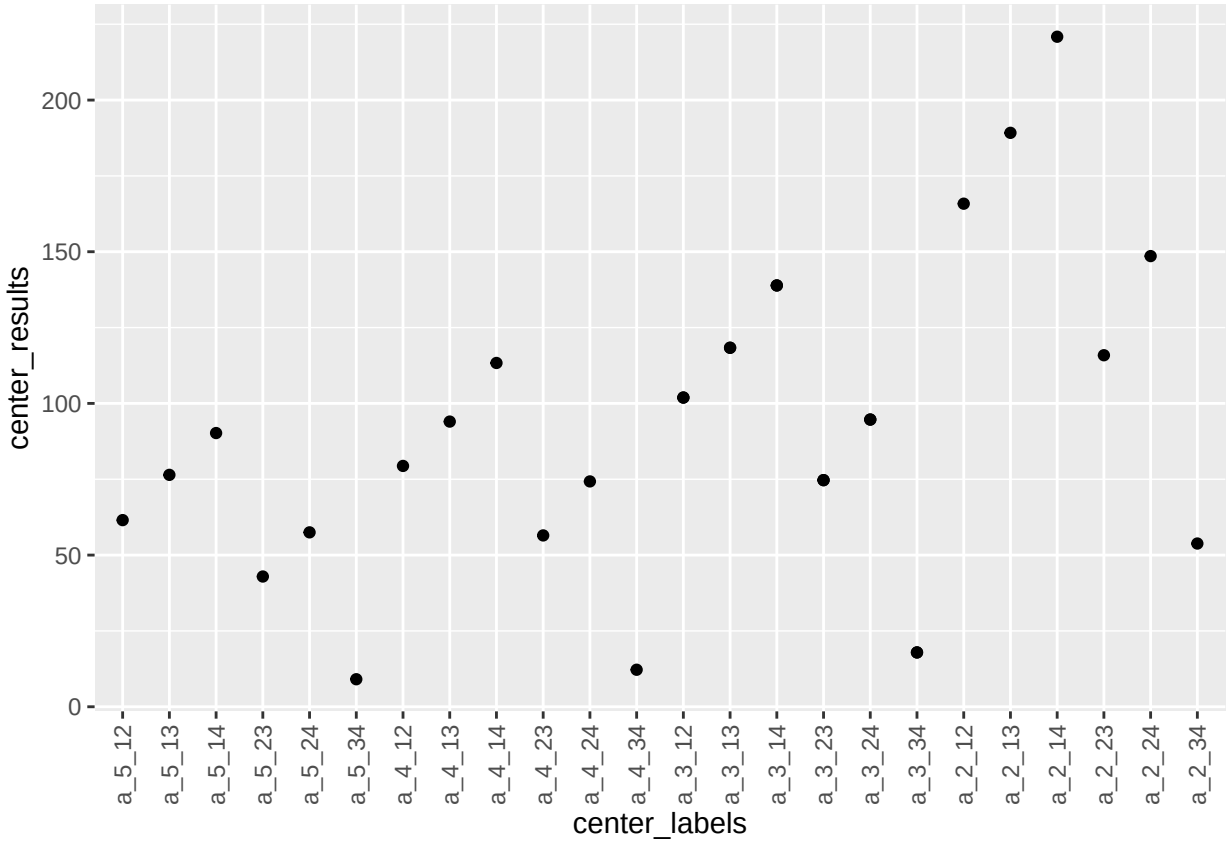
```
all_results <- arrange(all_results, center_results)

ggplot(
        all_results, aes(y = center_labels, x = center_results)) +
        geom_point() +
        theme(axis.text.x = element_text(angle = 90, vjust = 0.5)
      )
```




```
all_results <- arrange(all_results, center_labels)

ggplot(
  all_results, aes(x = center_labels, y = center_results)) +
  geom_point() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5)
  )
```


center_results
200
150
100
50
0
a_5_12
a_5_13
a_5_14
a_5_23
a_5_24
a_5_34
a_4_12
a_4_13
a_4_14
a_4_23
a_4_24
a_4_34
a_3_12
a_3_13
a_3_14
a_3_23
a_3_24
a_3_34
a_2_12
a_2_13
a_2_14
a_2_23
a_2_24
a_2_34
center_labels