stove - clustering

2022.08.24.

Table of contents

1 Introduction

1) 본 문서는 stove 패키지를 Shiny app에서 사용하는 것을 상정해 작성했습니다.

2) 본 문서의 케이스 스타일은 Camel case와 Snake case가 혼용되어 있습니다.

- Camel case : stove의 함수명 및 파라미터명
- Snake case: 유저로부터 받는 입력, shiny app의 server에서 사용(될 것이라고 예상)하는 object명, snake case로 작성된 dependencies의 함수명 등

2 Import sample data

1) 전처리가 완료된 샘플데이터를 불러옵니다.

- NA가 없어야 함
- string value가 있는 열은 factor로 변환

- 한 열이 모두 같은 값으로 채워져 있을 경우 제외해야 함
- Date type column이 없어야 함
- Outcome 변수는 classification의 경우 factor, regression의 경우 numeric이어야 함 (clustering은 outcome변수를 사용하지 않음)

```
# remotes::install_github("statgarten/datatoys")
library(stove)
library(datatoys)
library(dplyr)

set.seed(1234)

cleaned_data <- datatoys::bloodTest

cleaned_data <- cleaned_data %>%
  sample_n(1000) %>%
  subset(select = -c(TG))
```

3 K-means clustering

```
# user input
max_k <- 15 # k = 2:max_k, <= number of columns
n_start <- 25 # attempts 25 initial configurations, <= 175
iter_max <- 10 # <= 5000
n_boot <- 100 # Used only for determining the number of clusters using gap statistic
algorithm = "Hartigan-Wong" ## "Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"
select_optimal <- "silhouette" # "silhouette", "gap_stat" // there's no mathematical definiti
seed <- 6471

# K-means clustering
km_model <- stove::kMeansClustering(data = cleaned_data,
                                    maxK = max_k,
                                    nStart = n_start,
                                    iterMax = iter_max,
                                    nBoot = n_boot,
                                    algorithm = algorithm,
                                    selectOptimal = select_optimal,
                                    seedNum = seed
                                    )
```

```
km_model$result
```

```
K-means clustering with 2 clusters of sizes 426, 574

Cluster means:
       SEX    AGE_G      HGB    TCHOL     HDL        ANE        IHD        STK
1 1.497653 14.98592 14.18991 231.4671 56.5493 0.05868545 0.06338028 0.05399061
2 1.480836 14.17596 13.91446 170.3484 52.4007 0.09059233 0.06620209 0.09407666

Clustering vector:
   [1] 1 1 1 1 2 2 2 2 1 2 1 2 1 2 1 1 2 1 2 2 1 2 2 1 1 1 2 2 2 1 1 1 2 2 1 2 1
  [38] 1 1 2 2 2 2 1 2 2 1 2 1 2 1 1 2 2 1 2 1 1 2 2 1 2 2 2 1 2 2 2 2 2 2 2 1 2
  [75] 2 2 2 2 2 2 2 1 2 2 1 1 2 2 1 1 1 2 1 1 2 2 1 1 1 1 2 2 2 1 1 1 2 2 1 1 1
 [112] 1 1 1 2 1 2 2 1 1 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2 2 2 2 2 2 1 2 1 2 1 2 1 2 2
 [149] 1 1 2 1 1 1 2 2 2 1 2 1 2 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 1 2 2 1 2
 [186] 2 2 2 2 1 2 1 1 1 1 1 1 2 1 2 2 2 1 2 1 1 1 1 2 2 2 2 2 2 1 2 1 1 2 2 1 1
 [223] 2 2 2 1 2 2 2 2 2 1 2 2 2 2 2 2 2 1 2 2 2 1 2 2 1 2 1 1 2 1 2 2 1 2 1 1 1 2
 [260] 2 1 2 2 2 1 2 1 2 2 2 1 1 1 2 2 1 2 1 2 2 1 1 2 2 1 2 1 2 2 1 2 2 2 2 2 2
 [297] 2 2 2 2 1 1 2 2 2 1 2 2 2 1 2 1 2 1 2 1 2 2 2 1 1 2 2 2 2 1 2 2 2 2 2 1 2
 [334] 2 1 1 1 1 2 1 1 2 1 1 2 2 2 1 2 2 2 2 1 1 2 1 2 2 2 2 2 1 2 2 1 2 1 1 2 2
 [371] 1 2 2 1 1 2 2 1 2 1 1 1 2 2 2 2 2 2 2 1 2 2 2 1 1 2 2 2 2 2 1 1 1 2 1 2 1
 [408] 1 2 1 2 1 2 2 2 1 2 2 1 1 1 2 1 1 1 1 1 1 2 1 1 2 2 2 1 1 1 2 2 1 2 1 2 1
 [445] 2 1 1 2 2 2 2 2 2 1 1 2 2 2 2 2 2 1 1 2 2 1 1 2 2 2 2 2 1 1 2 1 2 2 2 1 2 2
 [482] 2 2 2 2 2 2 2 2 1 2 2 2 2 2 1 1 2 2 2 2 2 1 2 2 1 1 1 2 2 2 1 1 1 2 2 1 2 2 2
 [519] 2 2 2 1 2 1 1 2 1 1 1 2 2 2 1 2 2 2 2 2 2 2 2 1 1 1 1 2 2 2 2 2 1 2 1 2 2 2
 [556] 1 2 2 1 1 2 1 1 1 2 2 2 2 2 2 1 2 1 2 2 2 2 1 2 1 1 1 2 2 2 1 2 2 1 1 1 1
 [593] 1 2 2 2 2 1 2 2 2 1 1 1 1 2 2 2 2 2 1 2 2 1 1 1 2 1 2 2 2 2 2 2 2 2 2 2 1
 [630] 2 2 1 2 1 2 1 2 2 2 1 1 1 2 1 1 1 2 1 1 1 2 1 2 2 2 2 1 1 2 1 2 1 2 2 2 1
 [667] 1 2 1 2 1 1 2 2 1 1 1 2 2 1 2 1 1 1 1 1 1 2 2 2 1 2 2 1 2 2 1 1 1 2 2 2 2
 [704] 1 2 2 1 2 2 1 2 1 2 1 2 1 1 2 1 2 2 1 1 2 2 1 2 2 2 1 2 2 2 1 2 1 1 2 2 2
 [741] 1 1 1 2 1 1 1 1 1 2 2 2 2 2 2 1 1 2 1 1 1 2 1 1 2 2 1 2 1 2 2 1 2 2 1 1 1
 [778] 2 2 1 1 2 1 1 1 1 2 2 1 2 2 2 1 2 1 1 1 2 1 1 2 2 1 2 1 2 1 2 2 2 1 1 1 2
 [815] 1 1 1 2 1 1 2 1 2 2 2 1 1 2 2 2 1 2 2 1 2 1 2 2 1 2 1 2 1 1 1 2 1 2 2 2 1
 [852] 1 1 2 2 2 1 1 1 2 2 2 2 1 1 1 1 2 1 2 2 1 1 1 2 2 1 1 2 2 2 2 1 1 1 2 1 2
 [889] 2 1 1 2 1 2 2 1 2 1 2 2 2 1 1 2 1 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2 1 1 1 1 2
 [926] 2 1 1 2 2 1 1 2 2 2 2 1 1 2 2 2 1 2 2 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2
 [963] 2 1 1 2 2 1 2 1 2 2 1 2 2 1 1 2 2 1 2 1 2 1 2 2 2 2 1 2 1 1 2 2 1 2 2 1 2
[1000] 1

Within cluster sum of squares by cluster:
[1] 391527.4 394370.1
```

```
 (between_SS / total_SS =  53.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

4 K-means clustering without hyperparameters

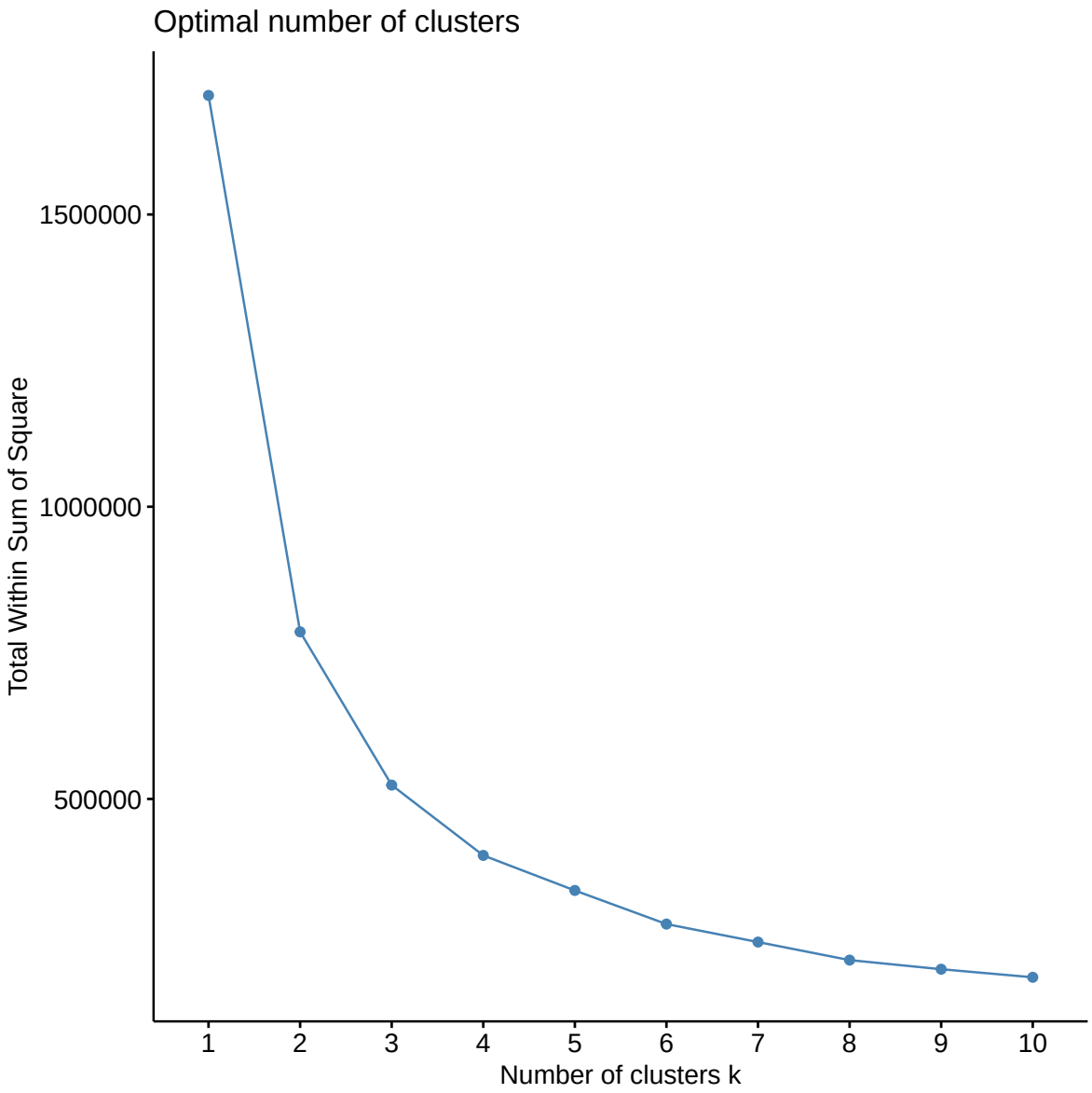
```r
# K-means clustering
km_model <- stove::kMeansClustering(data = cleaned_data)

km_model$result
```

```
K-means clustering with 2 clusters of sizes 426, 574

Cluster means:
       SEX    AGE_G      HGB    TCHOL     HDL        ANE        IHD        STK
1 1.497653 14.98592 14.18991 231.4671 56.5493 0.05868545 0.06338028 0.05399061
2 1.480836 14.17596 13.91446 170.3484 52.4007 0.09059233 0.06620209 0.09407666

Clustering vector:
   [1] 1 1 1 1 2 2 2 2 1 2 1 2 1 2 1 1 2 1 2 2 1 2 2 1 1 1 2 2 2 1 1 1 2 2 1 2 1
  [38] 1 1 2 2 2 2 1 2 2 1 2 1 2 1 1 2 2 1 2 1 1 2 2 1 2 2 2 1 2 2 2 2 2 2 2 1 2
  [75] 2 2 2 2 2 2 2 1 2 2 1 1 2 2 1 1 1 2 1 1 2 2 1 1 1 1 2 2 2 1 1 1 2 2 1 1 1
 [112] 1 1 1 2 1 2 2 1 1 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2 2 2 2 2 1 2 1 2 1 2 1 2 2
 [149] 1 1 2 1 1 1 2 2 2 1 2 1 2 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 1 2 2 1 2
 [186] 2 2 2 2 1 2 1 1 1 1 1 1 2 1 2 2 2 1 2 1 1 1 1 2 2 2 2 2 2 1 2 1 1 2 2 1 1
 [223] 2 2 2 1 2 2 2 2 1 2 2 2 2 2 2 2 1 2 2 2 1 2 2 1 2 1 1 2 1 2 2 1 2 1 1 1 2
 [260] 2 1 2 2 2 1 2 1 2 2 2 1 1 1 2 2 1 2 1 2 2 1 1 2 2 1 2 1 2 2 1 2 2 2 2 2 2
 [297] 2 2 2 2 1 1 2 2 2 1 2 2 2 1 2 1 2 1 2 1 2 2 2 1 1 2 2 2 2 1 2 2 2 2 2 1 2
 [334] 2 1 1 1 1 2 1 1 2 1 1 2 2 2 1 2 2 2 2 1 1 2 1 2 2 2 2 2 1 2 2 1 2 1 1 2 2
 [371] 1 2 2 1 1 2 2 1 2 1 1 1 2 2 2 2 2 2 2 1 2 2 2 1 1 2 2 2 2 2 1 1 1 2 1 2 1
 [408] 1 2 1 2 1 2 2 2 1 2 2 1 1 1 2 1 1 1 1 1 1 2 1 1 2 2 2 1 1 1 2 2 1 2 1 2 1
 [445] 2 1 1 2 2 2 2 2 2 2 1 1 2 2 2 2 2 2 1 1 2 2 1 1 2 2 2 2 2 1 1 2 1 2 2 2 1
 [482] 2 2 2 2 2 2 2 2 1 2 2 2 2 2 1 1 2 2 2 2 1 2 2 1 1 1 2 2 2 1 1 1 2 2 1 2 2
 [519] 2 2 2 1 2 1 1 2 1 1 1 2 2 2 1 2 2 2 2 2 2 2 1 1 1 1 2 2 2 2 2 1 2 1 2 2 2
 [556] 1 2 2 1 1 2 1 1 1 2 2 2 2 2 2 1 2 1 2 2 2 2 1 2 1 1 1 2 2 2 1 2 2 1 1 1 1
 [593] 1 2 2 2 2 1 2 2 2 1 1 1 1 2 2 2 2 2 1 2 2 1 1 1 2 1 2 2 2 2 2 2 2 2 2 2 1
```

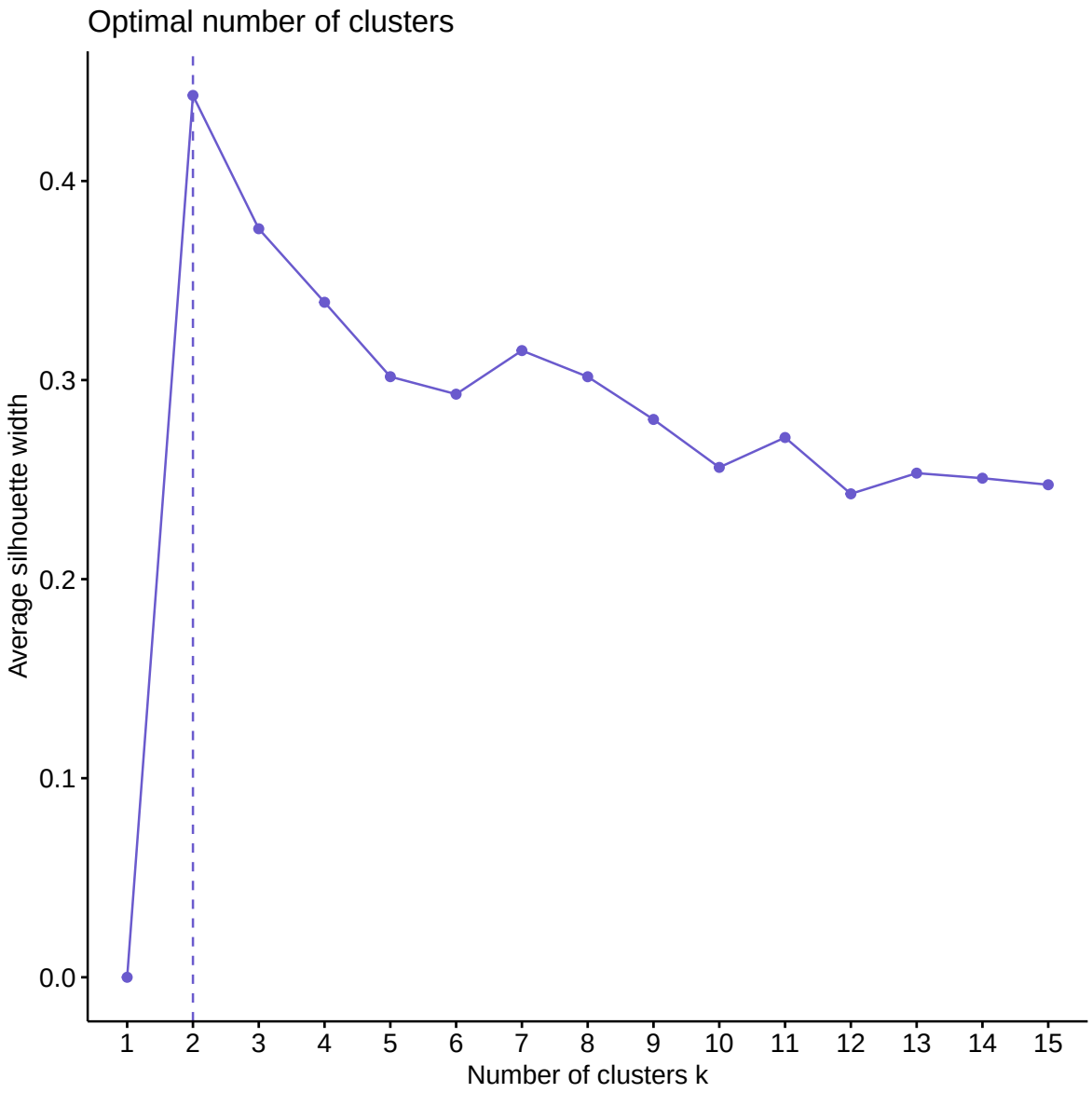
```
 [630] 2 2 1 2 1 2 1 2 2 2 1 1 1 2 1 1 1 2 1 1 1 2 1 2 2 2 2 1 1 2 1 2 1 2 2 2 1
 [667] 1 2 1 2 1 1 2 2 1 1 1 2 2 1 2 1 1 1 1 1 1 2 2 2 1 2 2 1 2 2 1 1 1 2 2 2 2
 [704] 1 2 2 1 2 2 1 2 1 2 1 2 1 1 2 1 2 2 1 1 2 2 1 2 2 2 1 2 2 2 1 2 1 1 2 2 2
 [741] 1 1 1 2 1 1 1 1 1 2 2 2 2 2 2 1 1 2 1 1 1 2 1 1 2 2 1 2 1 2 2 1 2 2 1 1 1
 [778] 2 2 1 1 2 1 1 1 1 2 2 1 2 2 2 1 2 1 1 1 2 1 1 2 2 1 2 1 2 1 2 2 2 1 1 1 2
 [815] 1 1 1 2 1 1 2 1 2 2 2 1 1 2 2 2 1 2 2 1 2 1 2 2 1 2 1 2 1 1 1 2 1 2 2 2 1
 [852] 1 1 2 2 2 1 1 1 2 2 2 2 1 1 1 1 2 1 2 2 1 1 1 2 2 1 1 2 2 2 2 1 1 1 2 1 2
 [889] 2 1 1 2 1 2 2 1 2 1 2 2 2 1 1 2 1 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2 1 1 1 1 2
 [926] 2 1 1 2 2 1 1 2 2 2 2 1 1 2 2 2 1 2 2 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 2
 [963] 2 1 1 2 2 1 2 1 2 2 1 2 2 1 1 2 2 1 2 1 2 1 2 2 2 2 1 2 1 1 2 2 1 2 2 1 2
[1000] 1

Within cluster sum of squares by cluster:
[1] 391527.4 394370.1
 (between_SS / total_SS =  53.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

5 Visualize

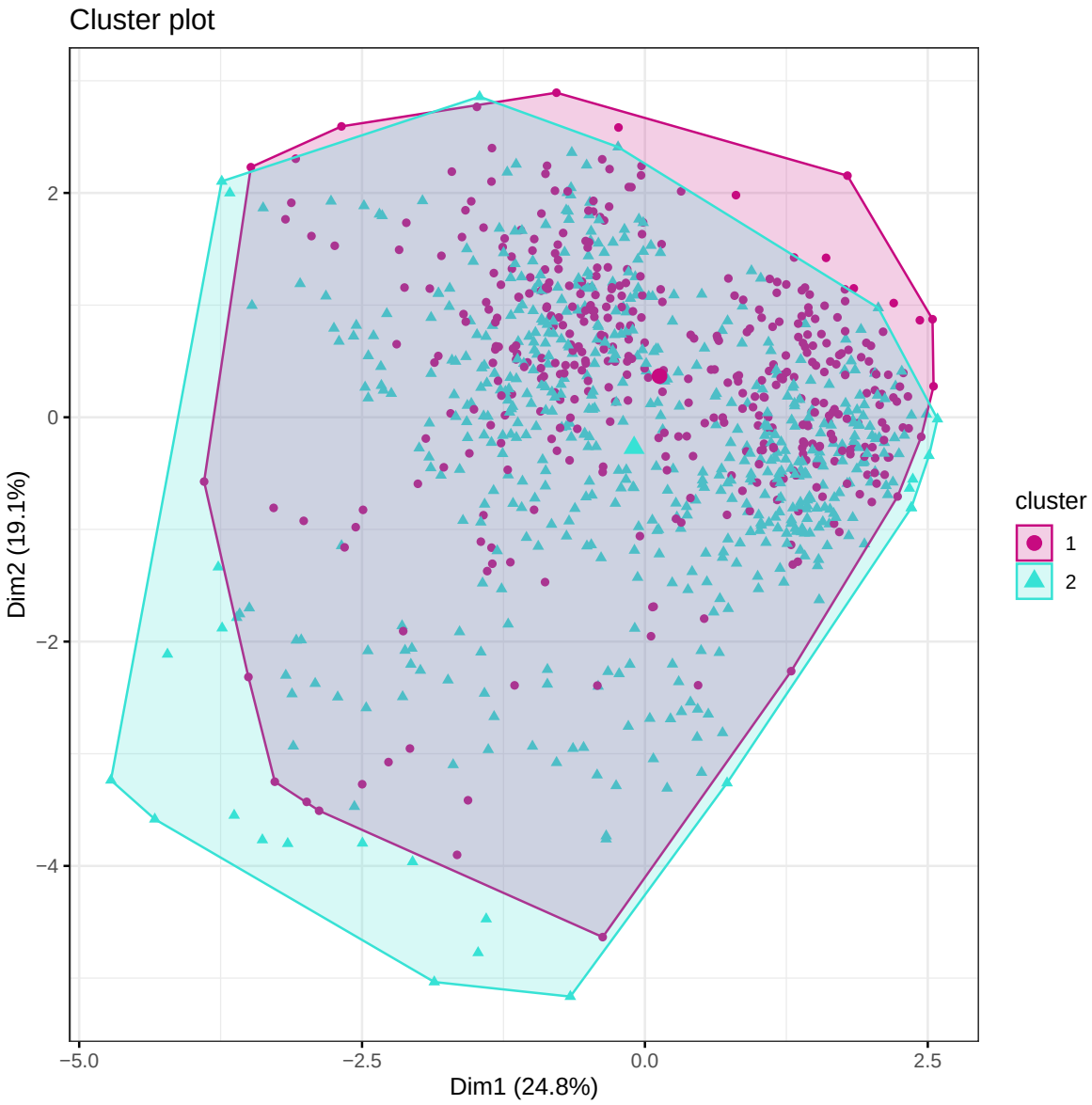
```
km_model$elbowPlot
```




```
km_model$optimalK
```




```
km_model$clustVis
```


Cluster plot
Dim2 (19.1%)
2
0
−2
−4
−5.0
−2.5
0.0
2.5
Dim1 (24.8%)
cluster
1
2