Model 3

Ying Luo

2022-12-16

K-Means

Helper packages

```
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
```

```
library(ggplot2)
library(stringr)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
# Modeling packages
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v tibble  3.1.8     v purrr   0.3.4
## v tidyr   1.2.1     v forcats 0.5.2
## v readr   2.1.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x gridExtra::combine() masks dplyr::combine()
## x dplyr::filter()      masks stats::filter()
## x dplyr::lag()         masks stats::lag()
```

```
library(cluster)
library(factoextra)
```

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

Load the dataset

```
library(readr)
df = read.csv("radiomics_completedata.csv")
```

Investigate the statistics of the dataset Output would not be presented to save pages

```
summary(df)
```

Remove NA

```
df <- na.omit(df)
```

Separate the training data (features) and their labels in the dataset

```
x_train <- data.matrix(df[-2])
label <- df[2]
```

Standardize the training data

```
x_train <- scale(x_train)
```

Investigate the Standardized the data Output would not be presented to save pages

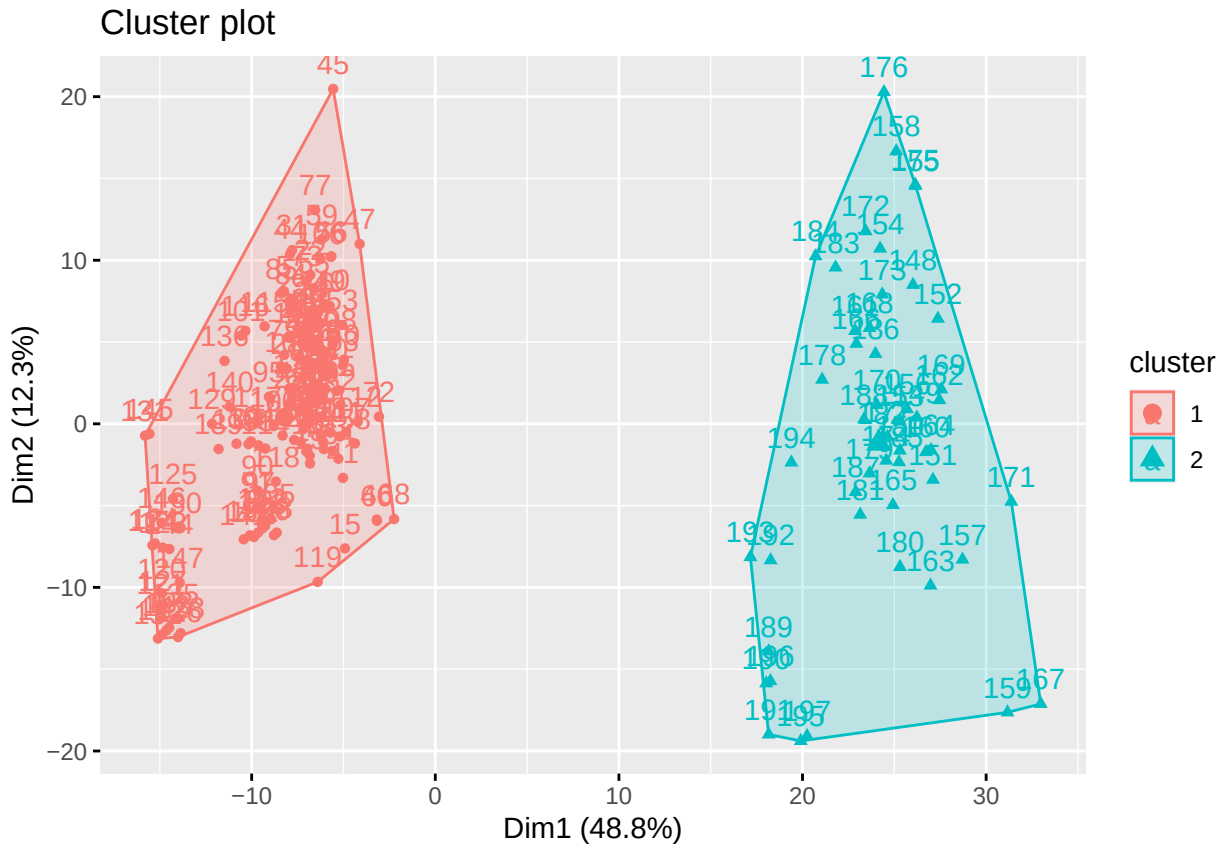
```
head(x_train)
```

Model building starts with k = 2 and result plotting

```
k2 <- kmeans(x_train, centers = 2, nstart = 25)
str(k2)
```

```
## List of 9
##  $ cluster     : Named int [1:197] 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "names")= chr [1:197] "1" "2" "3" "4" ...
##  $ centers     : num [1:2, 1:430] -0.537727 1.580918 0.000501 -0.001473 -0.016481 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "1" "2"
##   .. ..$ : chr [1:430] "Institution" "Failure" "Entropy_cooc.W.ADC" "GLNU_align.H.PET" ...
##  $ totss       : num 84280
##  $ withinss    : num [1:2] 23827 21069
##  $ tot.withinss: num 44895
##  $ betweenss   : num 39385
##  $ size        : int [1:2] 147 50
##  $ iter        : int 1
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
# Result plotting
fviz_cluster(k2, data = x_train)
```

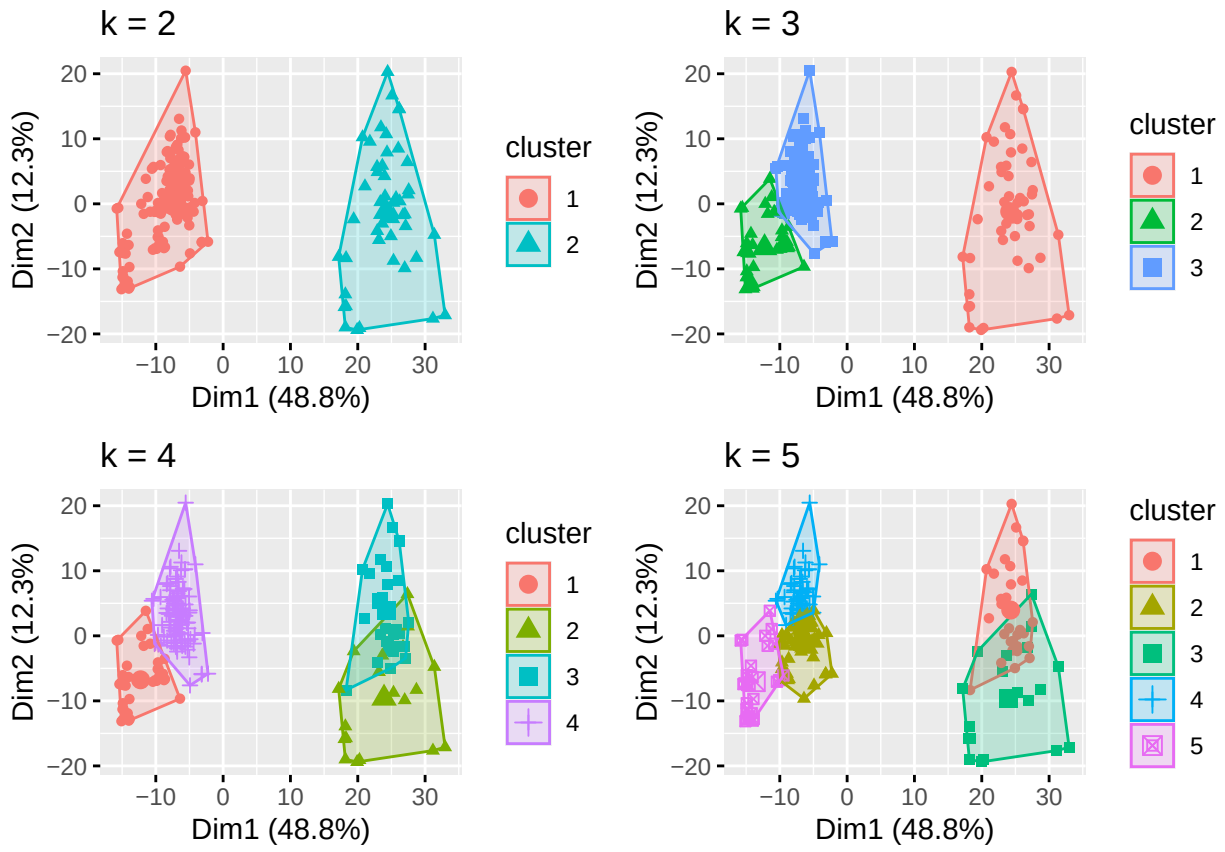


Model building experiments with different k values

```
k3 <- kmeans(x_train, centers = 3, nstart = 25)
k4 <- kmeans(x_train, centers = 4, nstart = 25)
k5 <- kmeans(x_train, centers = 5, nstart = 25)
```

Plot the results with different k values

```
p1 <- fviz_cluster(k2, geom = "point", data = x_train) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point",  data = x_train) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point",  data = x_train) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point",  data = x_train) + ggtitle("k = 5")

grid.arrange(p1, p2, p3, p4, nrow = 2)
```

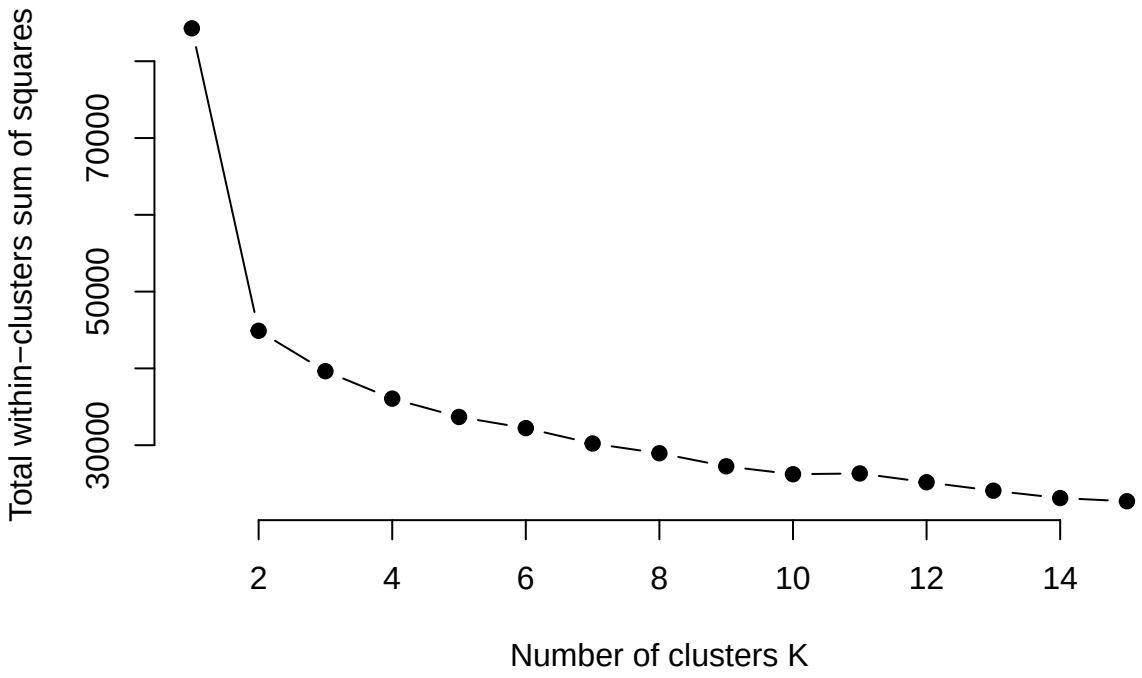


Finding the Optimal k value by computing total within-cluster sum of square

```
set.seed(123)
wss <- function(k) {
  kmeans(x_train, k, nstart = 10 )$tot.withinss
}
```

Compute and plot wss for k = 1 to k = 15

```
k.values <- 1:15
wss_values <- map_dbl(k.values, wss)

plot(k.values, wss_values,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```

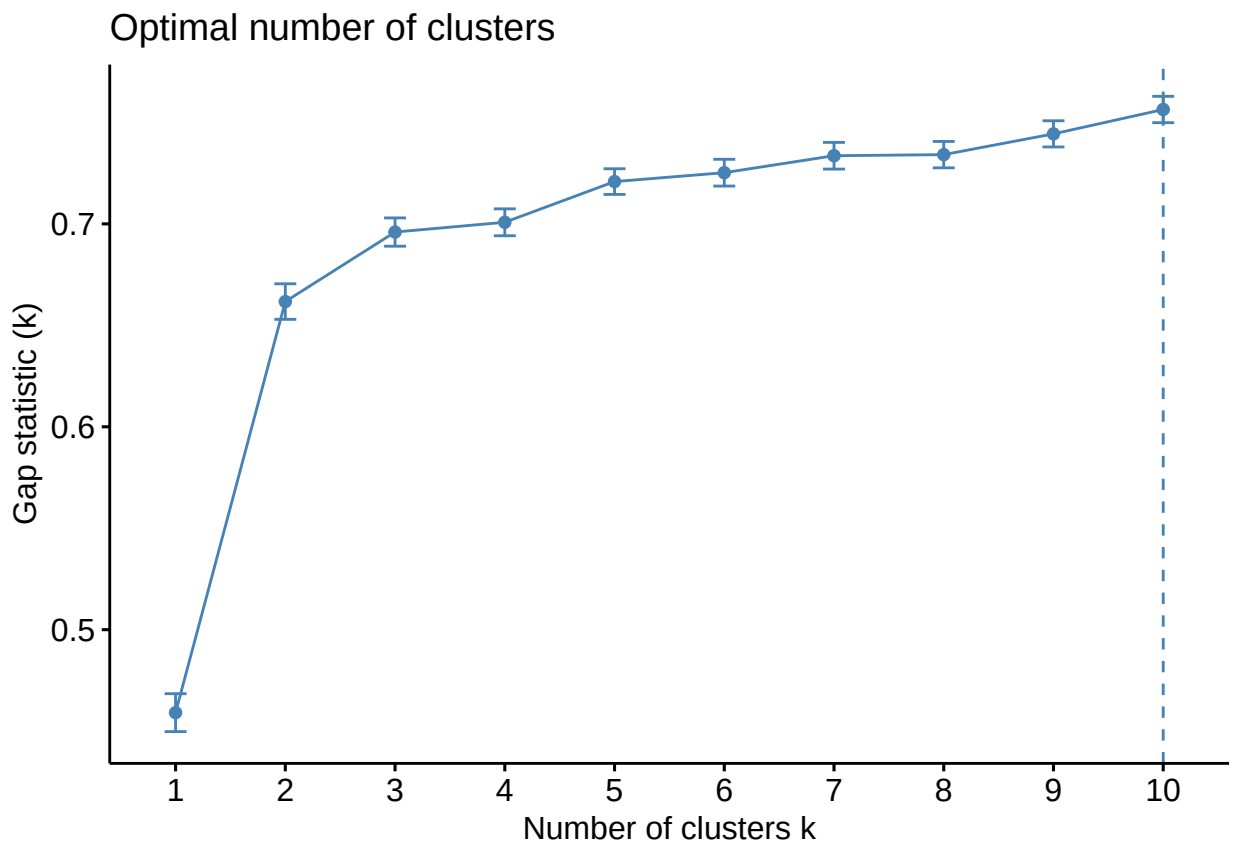


Compute gap statistic

```
set.seed(123)
gap_stat <- clusGap(x_train, FUN = kmeans, nstart = 25,
                    K.max = 10, B = 50)

print(gap_stat, method = "firstmax")
```

```
## Clustering Gap statistic ["clusGap"] from call:
## clusGap(x = x_train, FUNcluster = kmeans, K.max = 10, B = 50,     nstart = 25)
## B=50 simulated reference sets, k = 1..10; spaceH0="scaledPCA"
##  --> Number of clusters (method 'firstmax'): 10
##           logW   E.logW       gap      SE.sim
##  [1,] 7.172238 7.631343 0.4591050 0.009344722
##  [2,] 6.879860 7.541570 0.6617097 0.008762466
##  [3,] 6.799071 7.495045 0.6959738 0.006975238
##  [4,] 6.760250 7.461038 0.7007876 0.006630901
##  [5,] 6.715751 7.436595 0.7208444 0.006319998
##  [6,] 6.689678 7.414915 0.7252363 0.006598069
##  [7,] 6.661851 7.395440 0.7335893 0.006576344
##  [8,] 6.643375 7.377503 0.7341279 0.006496770
##  [9,] 6.616695 7.361058 0.7443634 0.006441963
## [10,] 6.589169 7.345537 0.7563681 0.006495757
```

```
fviz_gap_stat(gap_stat)
```



The final k value is determined to be 2 based on the above experiments and considering that the dataset has binary labels.

```
set.seed(123)
final <- kmeans(x_train, 2, nstart = 25)
print(final)
```

```
## K-means clustering with 2 clusters of sizes 50, 147
##
## Cluster means:
##   Institution       Failure Entropy_cooc.W.ADC GLNU_align.H.PET Min_hist.PET
## 1   1.5809179 -0.0014733768         0.04845450      -0.07901100    0.9204612
## 2  -0.5377272  0.0005011486        -0.01648112       0.02687449   -0.3130820
##   Max_hist.PET Mean_hist.PET Variance_hist.PET Standard_Deviation_hist.PET
## 1    0.9468341     0.9216792         0.4594337                   0.9319222
## 2   -0.3220524    -0.3134963        -0.1562700                  -0.3169804
##   Skewness_hist.PET Kurtosis_hist.PET Energy_hist.PET Entropy_hist.PET
## 1         0.9115602        0.25274217       0.6864958        1.5003007
## 2        -0.3100545       -0.08596673      -0.2335020       -0.5103064
##   AUC_hist.PET  H_suv.PET Volume.PET X3D_surface.PET ratio_3ds_vol.PET
## 1    1.6957546  0.9652219  0.5900077       0.3802612         0.9436984
## 2   -0.5767873 -0.3283068 -0.2006829      -0.1293406        -0.3209858
##   ratio_3ds_vol_norm.PET irregularity.PET tumor_length.PET Compactness_v1.PET
```

```
## 1              0.9622506        1.6522842        1.0256292          0.8807232
## 2             -0.3272961       -0.5620014       -0.3488535         -0.2995657
##   Compactness_v2.PET Spherical_disproportion.PET Sphericity.PET Asphericity.PET
## 1          0.4324058                   0.9622506      0.4460709       0.9240341
## 2         -0.1470768                  -0.3272961     -0.1517248      -0.3142973
##   Center_of_mass.PET Max_3D_diam.PET Major_axis_length.PET
## 1          0.6358358       0.8259982             0.8904297
## 2         -0.2162707      -0.2809518            -0.3028672
##   Minor_axis_length.PET Least_axis_length.PET Elongation.PET Flatness.PET
## 1             1.1433164             0.9772289      1.4563692    1.3553445
## 2            -0.3888831            -0.3323908     -0.4953637   -0.4610015
##   Max_cooc.L.PET Average_cooc.L.PET Variance_cooc.L.PET Entropy_cooc.L.PET
## 1      0.7290795           1.389215           1.1041050          1.6813985
## 2     -0.2479862          -0.472522          -0.3755459         -0.5719043
##   DAVE_cooc.L.PET DVAR_cooc.L.PET DENT_cooc.L.PET SAVE_cooc.L.PET
## 1       1.2936781       1.1366603       1.6603800       1.3889879
## 2      -0.4400266      -0.3866192      -0.5647551      -0.4724449
##   SVAR_cooc.L.PET SENT_cooc.L.PET ASM_cooc.L.PET Contrast_cooc.L.PET
## 1       1.1209781       1.6614758      0.6775498           0.9285775
## 2      -0.3812851      -0.5651278     -0.2304591          -0.3158427
##   Dissimilarity_cooc.L.PET Inv_diff_cooc.L.PET Inv_diff_norm_cooc.L.PET
## 1                1.2936781            1.443028                1.6979660
## 2               -0.4400266           -0.490826               -0.5775395
##   IDM_cooc.L.PET IDM_norm_cooc.L.PET Inv_var_cooc.L.PET Correlation_cooc.L.PET
## 1      1.2814891           1.7046571          1.2896785                1.123648
## 2     -0.4358807          -0.5798153         -0.4386661               -0.382193
##   Autocorrelation_cooc.L.PET Tendency_cooc.L.PET Shade_cooc.L.PET
## 1                  1.0338012           1.1209781        0.5578271
## 2                 -0.3516331          -0.3812851       -0.1897371
##   Prominence_cooc.L.PET IC1_.L.PET IC2_.L.PET Coarseness_vdif_.L.PET
## 1             0.7889007 -0.6341334  1.5273752              0.7537450
## 2            -0.2683336  0.2156916 -0.5195154             -0.2563758
##   Contrast_vdif_.L.PET Busyness_vdif_.L.PET Complexity_vdif_.L.PET
## 1            0.3878173            0.5565230              1.2153015
## 2           -0.1319107           -0.1892936             -0.4133678
##   Strength_vdif_.L.PET SRE_align.L.PET LRE_align.L.PET GLNU_align.L.PET
## 1            0.4934069        1.706523       1.6948229        0.4587983
## 2           -0.1678255       -0.580450      -0.5764704       -0.1560539
##   RLNU_align.L.PET RP_align.L.PET LGRE_align.L.PET HGRE_align.L.PET
## 1        0.4189336      1.7061400        1.0408063        1.0700373
## 2       -0.1424944     -0.5803197       -0.3540158       -0.3639583
##   LGSRE_align.L.PET HGSRE_align.L.PET LGHRE_align.L.PET HGLRE_align.L.PET
## 1          1.048281         1.0672364         1.0052958          1.078233
## 2         -0.356558        -0.3630056        -0.3419373         -0.366746
##   GLNU_norm_align.L.PET RLNU_norm_align.L.PET GLVAR_align.L.PET
## 1             1.1041018             1.7034139         1.1510468
## 2            -0.3755448            -0.5793925        -0.3915125
##   RLVAR_align.L.PET Entropy_align.L.PET SZSE.L.PET LZSE.L.PET LGLZE.L.PET
## 1         1.0474522           1.6880661  1.6676802  1.1852630   1.0601400
## 2        -0.3562762          -0.5741722 -0.5672382 -0.4031507  -0.3605919
##   HGLZE.L.PET SZLGE.L.PET SZHGE.L.PET LZLGE.L.PET LZHGE.L.PET GLNU_area.L.PET
## 1   1.0866745   1.0735299   1.0776043   0.8457163   0.8914749       0.4621309
## 2  -0.3696172  -0.3651462  -0.3665321  -0.2876586  -0.3032228      -0.1571874
##   ZSNU.L.PET ZSP.L.PET GLNU_norm.L.PET ZSNU_norm.L.PET GLVAR_area.L.PET
```

```
## 1  0.4218710  1.679008       1.1042309        1.681848        1.1694826
## 2 -0.1434935 -0.571091      -0.3755887       -0.572057       -0.3977832
##   ZSVAR.L.PET Entropy_area.L.PET Max_cooc.H.PET Average_cooc.H.PET
## 1   0.7548095          1.6893793      0.5052232          1.6652563
## 2  -0.2567379         -0.5746188     -0.1718446         -0.5664137
##   Variance_cooc.H.PET Entropy_cooc.H.PET DAVE_cooc.H.PET DVAR_cooc.H.PET
## 1           1.4721984          1.4404122       1.5079528       1.4645709
## 2          -0.5007478         -0.4899361      -0.5129091      -0.4981534
##   DENT_cooc.H.PET SAVE_cooc.H.PET SVAR_cooc.H.PET SENT_cooc.H.PET
## 1       1.3368883       1.6782221       1.4484331       1.1582831
## 2      -0.4547239      -0.5708239      -0.4926643      -0.3939739
##   ASM_cooc.H.PET Contrast_cooc.H.PET Dissimilarity_cooc.H.PET
## 1      0.4701159            1.344935                1.5079528
## 2     -0.1599034           -0.457461               -0.5129091
##   Inv_diff_cooc.H.PET Inv_diff_norm_cooc.H.PET IDM_cooc.H.PET
## 1           1.1377441                1.6996628      0.9576980
## 2          -0.3869878               -0.5781166     -0.3257476
##   IDM_norm_cooc.H.PET Inv_var_cooc_.H.PET Correlation_cooc.H.PET
## 1           1.7052806           0.9554037              1.1365587
## 2          -0.5800274          -0.3249672             -0.3865846
##   Autocorrelation_cooc.H.PET Tendency_cooc.H.PET Shade_cooc.H.PET
## 1                  1.5649714           1.4092944       -0.7124616
## 2                 -0.5323032          -0.4793518        0.2423339
##   Prominence_cooc.H.PET IC1_d.H.PET IC2_d.H.PET Coarseness_vdif.H.PET
## 1             1.0427158 -0.23095606   1.3345708             0.6663547
## 2            -0.3546653  0.07855648  -0.4539356            -0.2266512
##   Contrast_vdif.H.PET Busyness_vdif.H.PET Complexity_vdif.H.PET
## 1           0.4860224          0.25301766             1.0958360
## 2          -0.1653138         -0.08606043            -0.3727333
##   Strength_vdif.H.PET SRE_align.H.PET LRE_align.H.PET RLNU_align.H.PET
## 1          0.03112072       1.6638495       1.0890098        0.4166644
## 2         -0.01058528      -0.5659352      -0.3704115       -0.1417226
##   RP_align.H.PET LGRE_align.H.PET HGRE_align.H.PET LGSRE_align.H.PET
## 1      1.6436641        0.7082866        1.5743684         0.7040204
## 2     -0.5590694       -0.2409138       -0.5354994        -0.2394627
##   HGSRE_align.H.PET LGHRE_align.H.PET HGLRE_align.H.PET GLNU_norm_align.H.PET
## 1         1.6533952         0.7311054         0.7453460             0.8572435
## 2        -0.5623793        -0.2486753        -0.2535191            -0.2915794
##   RLNU_norm_align.H.PET GLVAR_align.H.PET RLVAR_align.H.PET Entropy_align.H.PET
## 1             1.5584253         1.4161797         0.4776867            1.550297
## 2            -0.5300766        -0.4816938        -0.1624785           -0.527312
##   SZSE.H.PET  LZSE.H.PET LGLZE.H.PET HGLZE.H.PET SZLGE.H.PET SZHGE.H.PET
## 1  1.4671263 -0.09759617   0.7096710   1.4890573   0.6984264   1.4294579
## 2 -0.4990226  0.03319598  -0.2413847  -0.5064821  -0.2375600  -0.4862102
##    LZLGE.H.PET LZHGE.H.PET GLNU_area.H.PET ZSNU.H.PET  ZSP.H.PET
## 1  0.001044652 -0.08592571       0.4835029  0.3648643  1.1565208
## 2 -0.000355324  0.02922643      -0.1644568 -0.1241035 -0.3933744
##   GLNU_norm.H.PET ZSNU_norm.H.PET GLVAR_area.H.PET ZSVAR_H.PET
## 1       0.8791603       1.2441418        1.3802703 -0.09449223
## 2      -0.2990341      -0.4231775       -0.4694797  0.03214021
##   Entropy_area.H.PET Max_cooc.W.PET Average_cooc.W.PET Variance_cooc.W.PET
## 1          1.6279234      0.5502762          0.9151412           0.4579807
## 2         -0.5537154     -0.1871688         -0.3112725          -0.1557757
##   Entropy_cooc.W.PET DAVE_cooc.W.PET DVAR_cooc.W.PET DENT_cooc.W.PET
```

```
## 1          1.4784780       0.9564701       0.5165571        1.450023
## 2         -0.5028837      -0.3253300      -0.1756997       -0.493205
##   SAVE_cooc.W.PET SVAR_cooc.W.PET SENT_cooc.W.PET ASM_cooc.W.PET
## 1       0.9140050       0.4135667       1.5336398      0.5955603
## 2      -0.3108861      -0.1406689      -0.5216462     -0.2025715
##   Contrast_cooc.W.PET Dissimilarity_cooc.W.PET Inv_diff_cooc.W.PET
## 1           0.5325478                0.9564701           1.2750883
## 2          -0.1811387               -0.3253300          -0.4337035
##   Inv_diff_norm_cooc.W.PET IDM_cooc.W.PET IDM_norm_cooc.W.PET
## 1                1.6983343       1.044167           1.7048157
## 2               -0.5776647      -0.355159          -0.5798693
##   Inv_var_cooc.W.PET Correlation_cooc.W.PET Autocorrelation_cooc.W.PET
## 1          1.1637708              1.1228422                  0.4576739
## 2         -0.3958404             -0.3819191                 -0.1556714
##   Tendency_cooc.W.PET Shade_cooc.W.PET Prominence_cooc.W.PET IC1_d.W.PET
## 1           0.4135667       0.07642004           0.022900737 -0.26887955
## 2          -0.1406689      -0.02599321          -0.007789366  0.09145563
##   IC2_d.W.PET Coarseness_vdif.W.PET Contrast_vdif.W.PET Busyness_vdif.W.PET
## 1   1.4455561             0.7071892           0.8252351           0.4153574
## 2  -0.4916858            -0.2405405          -0.2806922          -0.1412780
##   Complexity_vdif.W.PET Strength_vdif.W.PET SRE_align.W.PET LRE_align.W.PET
## 1             0.2991726           0.4249851        1.697315       1.4801473
## 2            -0.1017594          -0.1445527       -0.577318      -0.5034515
##   GLNU_align.W.PET RLNU_align.W.PET RP_align.W.PET LGRE_align.W.PET
## 1        0.4738278        0.4182280      1.6901986        0.8300003
## 2       -0.1611659       -0.1422544     -0.5748975       -0.2823130
##   HGRE_align.W.PET LGSRE_align.W.PET HGSRE_align.W.PET LGHRE_align.W.PET
## 1        0.4630749         0.8904857         0.4557129         0.5563026
## 2       -0.1575085        -0.3028863        -0.1550044        -0.1892186
##   HGLRE_align.W.PET GLNU_norm_align.W.PET RLNU_norm_align.W.PET
## 1         0.4921754             0.8494549              1.658483
## 2        -0.1674066            -0.2889302             -0.564110
##   GLVAR_align.W.PET RLVAR_align.W.PET Entropy_align.W.PET SZSE.W.PET
## 1         0.4593218         0.5957178           1.5543465  1.6121174
## 2        -0.1562319        -0.2026251          -0.5286893 -0.5483392
##    LZSE.W.PET LGLZE.W.PET HGLZE.W.PET SZLGE.W.PET SZHGE.W.PET  LZLGE.W.PET
## 1  0.21517025   0.8709408   0.4690713   0.9938480   0.4481637 -0.004326372
## 2 -0.07318716  -0.2962384  -0.1595481  -0.3380435  -0.1524366  0.001471555
##   LZHGE.W.PET GLNU_area.W.PET ZSNU.W.PET  ZSP.W.PET GLNU_norm.W.PET
## 1   0.5263985       0.4910918  0.3971868  1.4948131       0.8826796
## 2  -0.1790471      -0.1670380 -0.1350976 -0.5084398      -0.3002311
##   ZSNU_norm.W.PET GLVAR_area.W.PET ZSVAR.W.PET Entropy_area.W.PET Min_hist.ADC
## 1       1.4869647        0.4655759  0.06408427          1.6167770    0.5724098
## 2      -0.5057703       -0.1583592 -0.02179737         -0.5499242   -0.1946972
##   Max_hist.ADC Mean_hist.ADC Variance_hist.ADC Standard_Deviation_hist.ADC
## 1    1.5075750     1.4864908         0.7599395                    1.2359485
## 2   -0.5127806    -0.5056091        -0.2584828                   -0.4203906
##   Skewness_hist.ADC Kurtosis_hist.ADC Energy_hist.ADC Entropy_hist.ADC
## 1         0.3899909         0.4662845       0.7015053        1.6284344
## 2        -0.1326500        -0.1586002      -0.2386073       -0.5538893
##   AUC_hist.ADC Volume.ADC X3D_surface.ADC ratio_3ds_vol.ADC
## 1    1.6655300  0.5687484       0.7349831         1.1042095
## 2   -0.5665068 -0.1934518      -0.2499942        -0.3755815
##   ratio_3ds_vol_norm.ADC irregularity.ADC Compactness_v1.ADC Compactness_v2.ADC
```

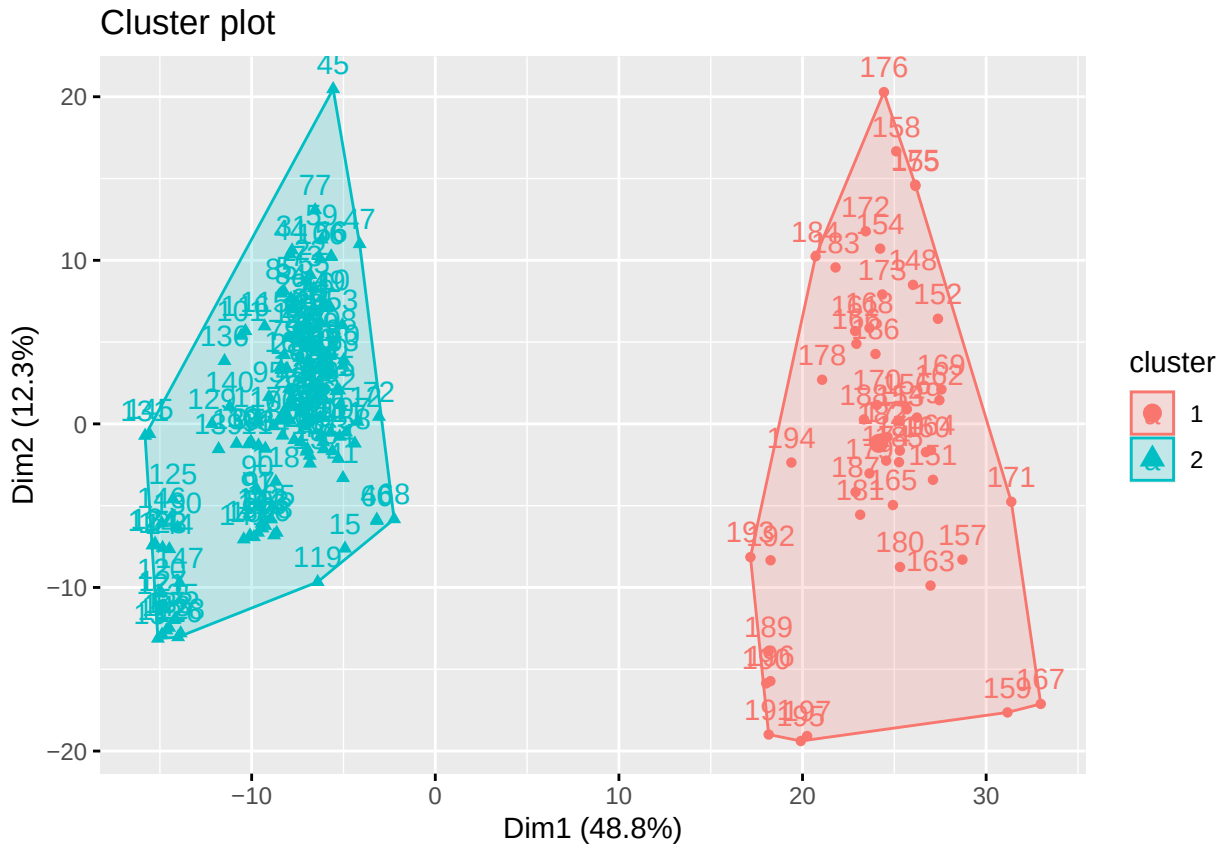
```
## 1              1.6106322        1.6397737          1.1221987          1.3007130
## 2             -0.5478341       -0.5577462         -0.3817002         -0.4424194
##   Spherical_disproportion.ADC Sphericity.ADC Asphericity.ADC Center_of_mass.ADC
## 1                   1.6106322      1.6242350       1.1989866          0.5373920
## 2                  -0.5478341     -0.5524609      -0.4078186         -0.1827864
##   Max_3D_diam.ADC Major_axis_length.ADC Minor_axis_length.ADC
## 1       1.0866100             1.2316275             1.1312333
## 2      -0.3695952            -0.4189209            -0.3847732
##   Least_axis_length.ADC Elongation.ADC Flatness.ADC Max_cooc.L.ADC
## 1             1.0417403      1.4824827    1.4052040      0.8250964
## 2            -0.3543334     -0.5042458   -0.4779606     -0.2806450
##   Average_cooc.L.ADC Variance_cooc.L.ADC Entropy_cooc.L.ADC DAVE_cooc.L.ADC
## 1           1.456079           0.9533869          1.6827114       1.2819538
## 2          -0.495265          -0.3242813         -0.5723508      -0.4360387
##   DVAR_cooc.L.ADC DENT_cooc.L.ADC SAVE_cooc.L.ADC SVAR_cooc.L.ADC
## 1       0.9295089       1.6521421       1.4558899       0.9317704
## 2      -0.3161595      -0.5619531      -0.4952006      -0.3169287
##   SENT_cooc.L.ADC ASM_cooc.L.ADC Contrast_cooc.L.ADC Dissimilarity_cooc.L.ADC
## 1       1.2584756      0.7127202           0.8811662                1.2819538
## 2      -0.4280529     -0.2424218          -0.2997164               -0.4360387
##   Inv_diff_cooc.L.ADC Inv_diff_norm_cooc.L.ADC IDM_cooc.L.ADC
## 1           1.5058302                1.7039344      1.3642322
## 2          -0.5121871               -0.5795695     -0.4640245
##   IDM_norm_cooc.L.ADC Inv_var_cooc.L.ADC Correlation_cooc.L.ADC
## 1           1.7073272           1.379898              1.2216811
## 2          -0.5807235          -0.469353             -0.4155378
##   Autocorrelation_.L.ADC Tendency_cooc.L.ADC Shade_.L.ADC Prominence_cooc.L.ADC
## 1              1.1050198           0.9317704   0.29259000             0.5515288
## 2             -0.3758571          -0.3169287  -0.09952041            -0.1875948
##   IC1_.L.ADC IC2_.L.ADC Coarseness_vdif_.L.ADC Contrast_vdif_.L.ADC
## 1 -0.6732168  1.5121032              0.6939723            0.6587722
## 2  0.2289853 -0.5143208             -0.2360450           -0.2240722
##   Busyness_vdif_.L.ADC Complexity_vdif_.L.ADC Strength_vdif_.L.ADC
## 1            0.6475886              1.2753146            0.4214397
## 2           -0.2202682             -0.4337805           -0.1433468
##   SRE_align.L.ADC LRE_align.L.ADC GLNU_align.L.ADC RLNU_align.L.ADC
## 1       1.7052408       1.6811893        0.5682374        0.5910147
## 2      -0.5800139      -0.5718331       -0.1932780       -0.2010254
##   RP_align.L.ADC LGRE_align.L.ADC HGRE_align.L.ADC LGSRE_align.L.ADC
## 1      1.7034645        0.7243458        1.2086645         0.7235521
## 2     -0.5794097       -0.2463761       -0.4111104        -0.2461061
##   HGSRE_align.L.ADC LGHRE_align.L.ADC HGLRE_align.L.ADC GLNU_norm_align.L.ADC
## 1         1.2124123         0.7234431         1.1801466             1.2291014
## 2        -0.4123852        -0.2460691        -0.4014104            -0.4180617
##   RLNU_norm_align.L.ADC GLVAR_align.L.ADC RLVAR_align.L.ADC Entropy_align.L.ADC
## 1             1.6955541         0.9930121         1.1385331           1.6982212
## 2            -0.5767191        -0.3377592        -0.3872562          -0.5776262
##   SZSE.L.ADC LZSE.L.ADC LGLZE.L.ADC HGLZE.L.ADC SZLGE.L.ADC SZHGE.L.ADC
## 1  1.6968578  1.3430968   0.7262967   1.2295659   0.7219542   1.2399482
## 2 -0.5771625 -0.4568356  -0.2470397  -0.4182197  -0.2455627  -0.4217511
##   LZLGE.L.ADC LZHGE.L.ADC GLNU_area.L.ADC ZSNU.L.ADC  ZSP.L.ADC GLNU_norm.L.ADC
## 1   0.6651854    1.077189       0.5782984  0.5919629  1.6748354       1.2251432
## 2  -0.2262535   -0.366391      -0.1967001 -0.2013479 -0.5696719      -0.4167154
##   ZSNU_norm.L.ADC GLVAR_area.L.ADC ZSVAR.L.ADC Entropy_area.L.ADC
```

```
## 1       1.6570978         1.012871   0.6758567          1.7010816
## 2      -0.5636387        -0.344514  -0.2298832         -0.5785992
##   Max_cooc.H.ADC Average_cooc.H.ADC Variance_cooc.H.ADC Entropy_cooc.H.ADC
## 1      0.7039103          1.6967547           1.7053247          1.7011475
## 2     -0.2394253         -0.5771274          -0.5800424         -0.5786216
##   DAVE_cooc.H.ADC DVAR_cooc.H.ADC DENT_cooc.H.ADC SAVE_cooc.H.ADC
## 1       1.5698813       1.4861394       1.7017575       1.6967573
## 2      -0.5339732      -0.5054896      -0.5788291      -0.5771283
##   SVAR_cooc.H.ADC SENT_cooc.H.ADC ASM_cooc.H.ADC Contrast_cooc.H.ADC
## 1       1.6206816       1.6803084      0.6607170           1.3858879
## 2      -0.5512522      -0.5715335     -0.2247337          -0.4713904
##   Dissimilarity_cooc.H.ADC Inv_diff_cooc.H.ADC Inv_diff_norm_cooc.H.ADC
## 1                1.5698813           1.5546888                1.7028145
## 2               -0.5339732          -0.5288057               -0.5791886
##   IDM_cooc.H.ADC IDM_norm_cooc.H.ADC Inv_var_cooc.H.ADC Correlation_cooc.H.ADC
## 1      1.4136874           1.7054539          1.4364367              1.1993586
## 2     -0.4808461          -0.5800864         -0.4885839             -0.4079451
##   Autocorrelation_cooc.H.ADC Tendency_cooc.H.ADC Shade_cooc.H.ADC
## 1                  1.6722184           1.6206816        0.3887230
## 2                 -0.5687818          -0.5512522       -0.1322187
##   Prominence_cooc.H.ADC IC1_d.H.ADC IC2_d.H.ADC Coarseness_vdif.H.ADC
## 1             1.5404751  -0.5455177   1.5085932             0.6780216
## 2            -0.5239711   0.1855502  -0.5131269            -0.2306196
##   Contrast_vdif.H.ADC Busyness_vdif.H.ADC Complexity_vdif.H.ADC
## 1           1.5316725           0.6153610              1.503704
## 2          -0.5209771          -0.2093065             -0.511464
##   Strength_vdif.H.ADC SRE_align.H.ADC LRE_align.H.ADC GLNU_align.H.ADC
## 1           0.3677298       1.7071497       1.7038845        0.5901231
## 2          -0.1250782      -0.5806632      -0.5795526       -0.2007222
##   RLNU_align.H.ADC RP_align.H.ADC LGRE_align.H.ADC HGRE_align.H.ADC
## 1        0.5924412       1.706814        1.0946139        1.7100780
## 2       -0.2015106      -0.580549       -0.3723177       -0.5816592
##   LGSRE_align.H.ADC HGSRE_align.H.ADC LGHRE_align.H.ADC HGLRE_align.H.ADC
## 1         1.0760014         1.7093907         1.1710039         1.7053139
## 2        -0.3659869        -0.5814254        -0.3983006        -0.5800387
##   GLNU_norm_align.H.ADC RLNU_norm_align.H.ADC GLVAR_align.H.ADC
## 1             0.9735389             1.7053279         1.7100152
## 2            -0.3311357            -0.5800435        -0.5816378
##   RLVAR_align.H.ADC Entropy_align.H.ADC SZSE.H.ADC LZSE.H.ADC LGLZE.H.ADC
## 1         1.0687509           1.7093530  1.7049082  1.6336887   1.0589022
## 2        -0.3635207          -0.5814126 -0.5799008 -0.5556764  -0.3601708
##   HGLZE.H.ADC SZLGE.H.ADC SZHGE.H.ADC LZLGE.H.ADC LZHGE.H.ADC GLNU_area.H.ADC
## 1    1.709075   1.0114862   1.7031396   1.0813161   1.5698347       0.5919958
## 2   -0.581318  -0.3440429  -0.5792992  -0.3677946  -0.5339574      -0.2013591
##   ZSNU.H.ADC  ZSP.H.ADC GLNU_norm.H.ADC ZSNU_norm.H.ADC GLVAR_area.H.ADC
## 1  0.5972096  1.7013318       0.9745507        1.692802        1.7072803
## 2 -0.2031325 -0.5786843      -0.3314798       -0.575783       -0.5807076
##   ZSVAR.H.ADC Entropy_area.H.ADC Max_cooc.W.ADC Average_cooc.W.ADC
## 1   0.8431301          1.7066118      0.6868122           1.199285
## 2  -0.2867790         -0.5804802     -0.2336096          -0.407920
##   Variance_cooc.W.ADC DAVE_cooc.W.ADC DVAR_cooc.W.ADC DENT_cooc.W.ADC
## 1           0.7283676       1.3033631       0.7679414       1.6768624
## 2          -0.2477441      -0.4433208      -0.2612045      -0.5703613
##   SAVE_cooc.W.ADC SVAR_cooc.W.ADC SENT_cooc.W.ADC ASM_cooc.W.ADC
```

```
## 1       1.1909017       0.6843706       1.2023295      0.6601442
## 2      -0.4050686      -0.2327791      -0.4089556     -0.2245389
##   Contrast_cooc.W.ADC Dissimilarity_cooc.W.ADC Inv_diff_cooc.W.ADC
## 1           0.7994120                1.3033631           1.3827605
## 2          -0.2719088               -0.4433208          -0.4703267
##   Inv_diff_norm_cooc.W.ADC IDM_cooc.W.ADC IDM_norm_cooc.W.ADC
## 1                1.7038802      1.3112119           1.7073083
## 2               -0.5795511     -0.4459904          -0.5807171
##   Inv_var_cooc.W.ADC Correlation_cooc.W.ADC Autocorrelation_cooc.W.ADC
## 1          1.3074526              1.2225367                  0.8447953
## 2         -0.4447118             -0.4158288                 -0.2873453
##   Tendency_cooc.W.ADC Shade_cooc.W.ADC Prominence_cooc.W.ADC IC1_d.W.ADC
## 1           0.6843706        0.2567335             0.3775512  -0.6756692
## 2          -0.2327791       -0.0873243            -0.1284188   0.2298194
##   IC2_d.W.ADC Coarseness_vdif.W.ADC Contrast_vdif.W.ADC Busyness_vdif.W.ADC
## 1   1.6012140             0.7114542           0.6249552           1.0116700
## 2  -0.5446306            -0.2419912          -0.2125698          -0.3441054
##   Complexity_vdif.W.ADC Strength_vdif.W.ADC SRE_align.W.ADC LRE_align.W.ADC
## 1             0.6003182           0.5784705       1.7073214       1.7065667
## 2            -0.2041899          -0.1967587      -0.5807216      -0.5804649
##   GLNU_align.W.ADC RLNU_align.W.ADC RP_align.W.ADC LGRE_align.W.ADC
## 1        0.6326468        0.5857336      1.7071535        0.6918953
## 2       -0.2151860       -0.1992291     -0.5806645       -0.2353386
##   HGRE_align.W.ADC LGSRE_align.W.ADC HGSRE_align.W.ADC LGHRE_align.W.ADC
## 1        0.8626770         0.6918084         0.8616174         0.6894568
## 2       -0.2934276        -0.2353090        -0.2930672        -0.2345091
##   HGLRE_align.W.ADC GLNU_norm_align.W.ADC RLNU_norm_align.W.ADC
## 1          0.866512             0.9154487             1.7063312
## 2         -0.294732            -0.3113771            -0.5803848
##   GLVAR_align.W.ADC RLVAR_align.W.ADC Entropy_align.W.ADC SZSE.W.ADC LZSE.W.ADC
## 1         0.7640782         0.9834635            1.661714  1.7066974  1.6823970
## 2        -0.2598905        -0.3345114           -0.565209 -0.5805093 -0.5722439
##   LGLZE.W.ADC HGLZE.W.ADC SZLGE.W.ADC SZHGE.W.ADC LZLGE.W.ADC LZHGE.W.ADC
## 1   0.6918923   0.8639228   0.6899145   0.8602645   0.6450074   0.8755515
## 2  -0.2353375  -0.2938513  -0.2346648  -0.2926070  -0.2193903  -0.2978066
##   GLNU_area.W.ADC ZSNU.W.ADC  ZSP.W.ADC GLNU_norm.W.ADC ZSNU_norm.W.ADC
## 1       0.6327545  0.5822861  1.7050925       0.9137899        1.699026
## 2      -0.2152226 -0.1980565 -0.5799634      -0.3108129       -0.577900
##   GLVAR_area.W.ADC ZSVAR.W.ADC Entropy_area.W.ADC
## 1        0.7713592   1.0785430           1.672228
## 2       -0.2623671  -0.3668514          -0.568785
##
## Clustering vector:
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
##  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
```

```
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
##   2   2   2   2   2   2   2   1   1   1   1   1   1   1   1   1   1   1   1   1
## 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##
## Within cluster sum of squares by cluster:
## [1] 21068.8 23826.7
##  (between_SS / total_SS =  46.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
fviz_cluster(final, data = x_train)
```



Hierarchical

Helper packages

```
library(dplyr)
library(ggplot2)

# Modeling packages
library(cluster)
library(factoextra)
```

Compute euclidean distance

```
set.seed(123)
distance <- dist(x_train, method = "euclidean")
```

Hierarchical clustering using Complete Linkage

```
hc1 <- hclust(distance, method = "complete" )
```

Compute complete linkage clustering with agnes and print the Agglomerative coefficient

```
set.seed(123)
hc2 <- agnes(x_train, method = "complete")

# Agglomerative coefficient
hc2$ac
```

```
## [1] 0.8488437
```

Different methods to evaluate

```
m <- c( "average", "single", "complete", "ward")
names(m) <- c( "average", "single", "complete", "ward")
```

Create function to compute coefficient and obtain the coefficient for each linkage method

```
ac <- function(x) {
  agnes(x_train, method = x)$ac
}

# get agglomerative coefficient for each linkage method
purrr::map_dbl(m, ac)
```

```
##   average    single  complete      ward
## 0.7618315 0.7097208 0.8488437 0.9655196
```

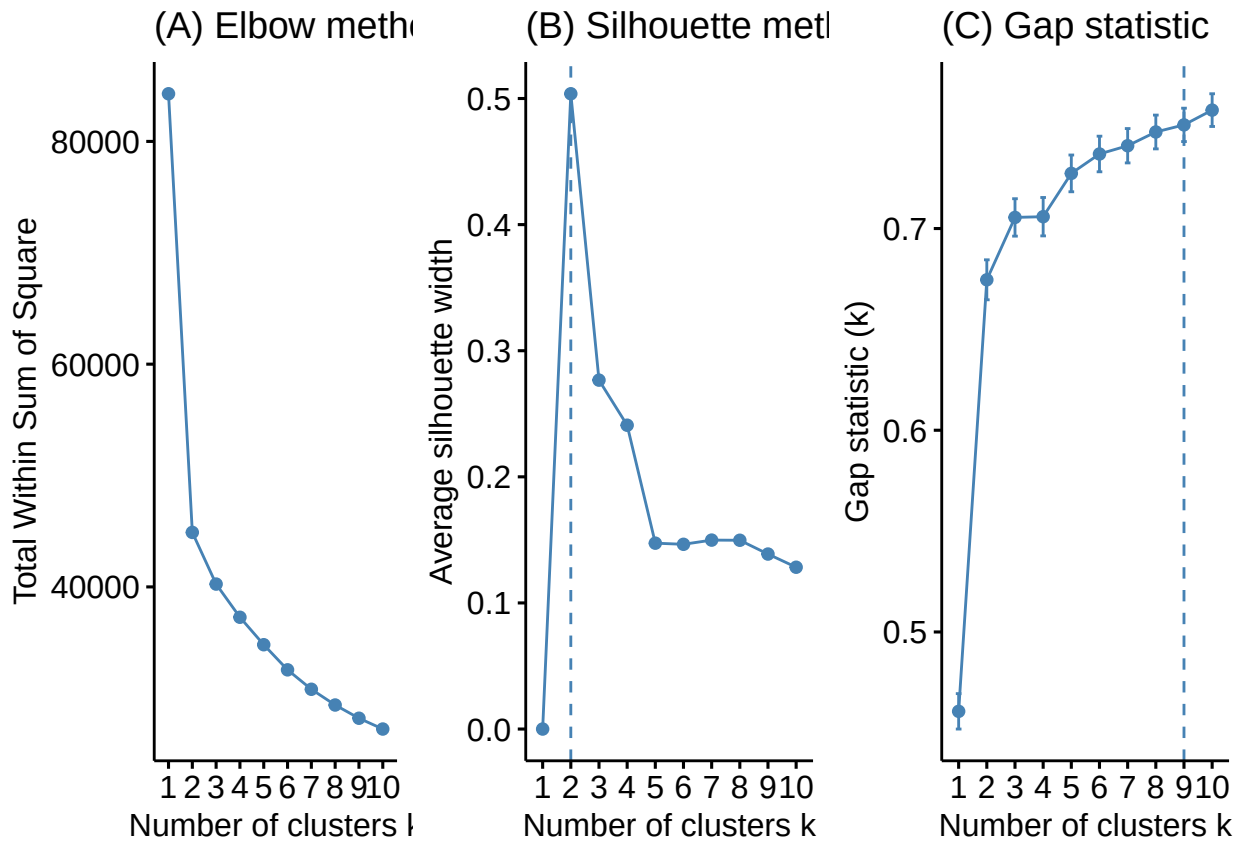
Compute divisive hierarchical clustering and print the Divise coefficient

```
hc3 <- diana(x_train)

# Divise coefficient; amount of clustering structure found
hc3$dc
```

```
## [1] 0.8427741
```

Plot cluster results

```
p1 <- fviz_nbclust(x_train, FUN = hcut, method = "wss",
                   k.max = 10) +
  ggtitle("(A) Elbow method")
p2 <- fviz_nbclust(x_train, FUN = hcut, method = "silhouette",
                   k.max = 10) +
  ggtitle("(B) Silhouette method")
p3 <- fviz_nbclust(x_train, FUN = hcut, method = "gap_stat",
                   k.max = 10) +
  ggtitle("(C) Gap statistic")

gridExtra::grid.arrange(p1, p2, p3, nrow = 1)
```



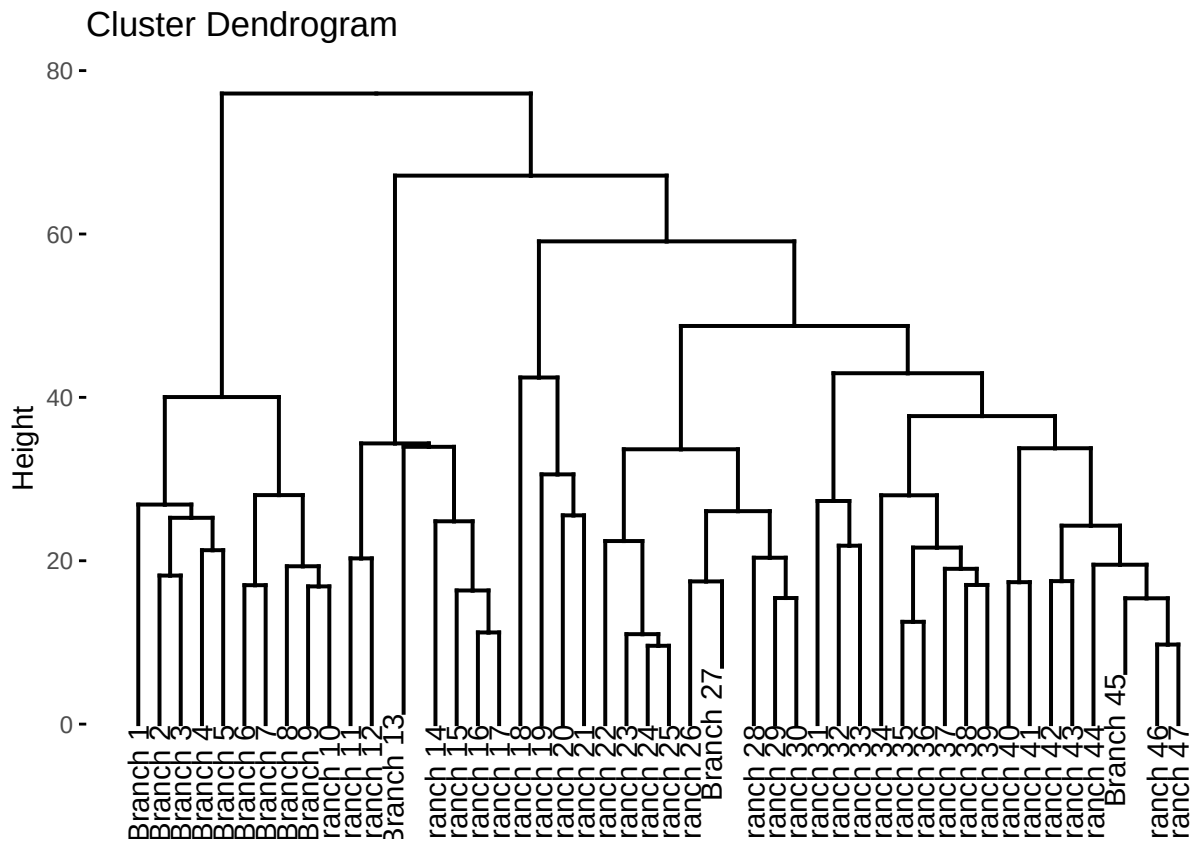
Construct dendorgram

```
hc4 <- hclust(distance, method = "ward.D2" )
dend_plot <- fviz_dend(hc4)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

```
dend_data <- attr(dend_plot, "dendrogram")
dend_cuts <- cut(dend_data, h = 8)
fviz_dend(dend_cuts$upper[[1]])
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```



Hierarchical clustering using using Ward's method

```
hc4 <- hclust(distance, method = "ward.D2" )
```

Cut tree into 4 groups

```
sub_grp <- cutree(hc4, k = 8)
```

Number of members in each cluster

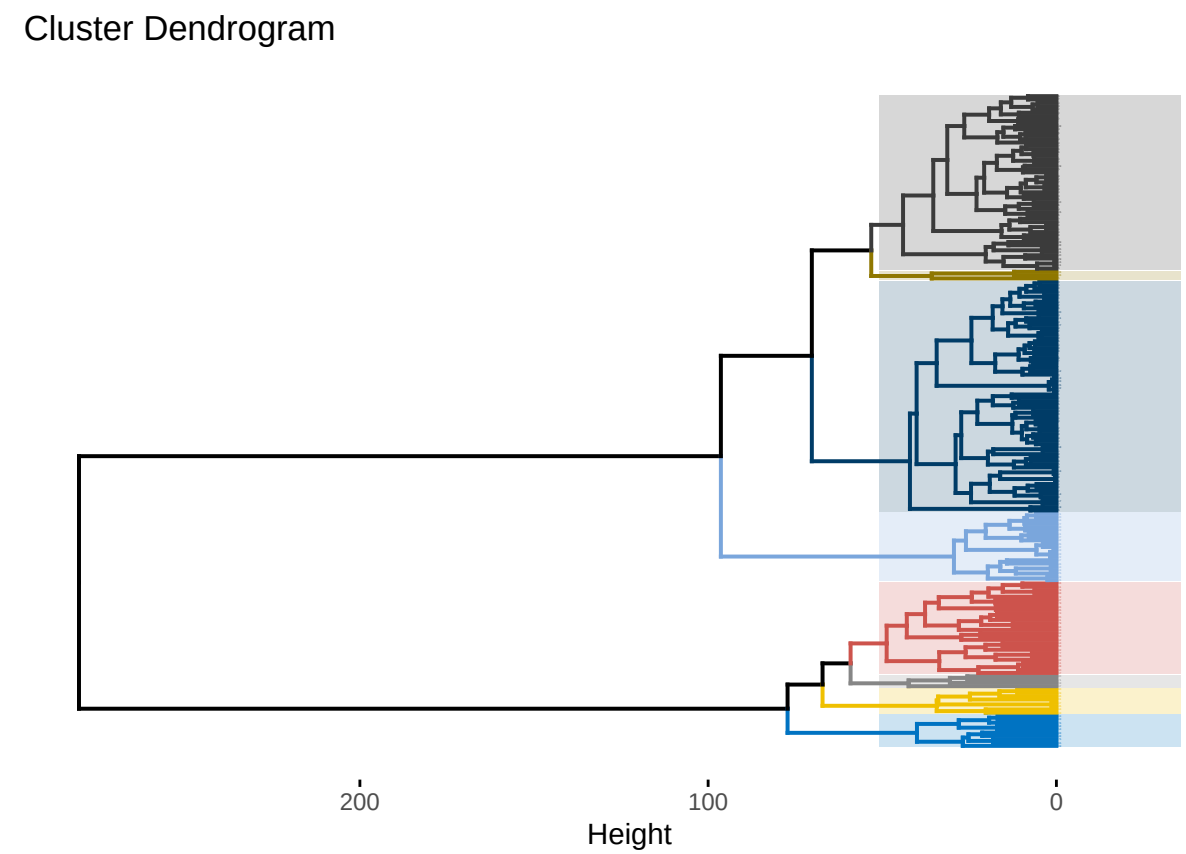
```
table(sub_grp)
```

```
## sub_grp
##  1  2  3  4  5  6  7  8
## 70 53  3 21 10 28  4  8
```

Plot the full dendogram

```
fviz_dend(
  hc4,
  k = 8,
  horiz = TRUE,
  rect = TRUE,
  rect_fill = TRUE,
  rect_border = "jco",
  k_colors = "jco",
  cex = 0.1
)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```




```
# create full dendogram
dend_plot <- fviz_dend(hc4)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

```
# extract plot info
dend_data <- attr(dend_plot, "dendrogram")

# cut the dendogram
dend_cuts <- cut(dend_data, h = 70.5)
```

Designated height Create sub dendrogram plots

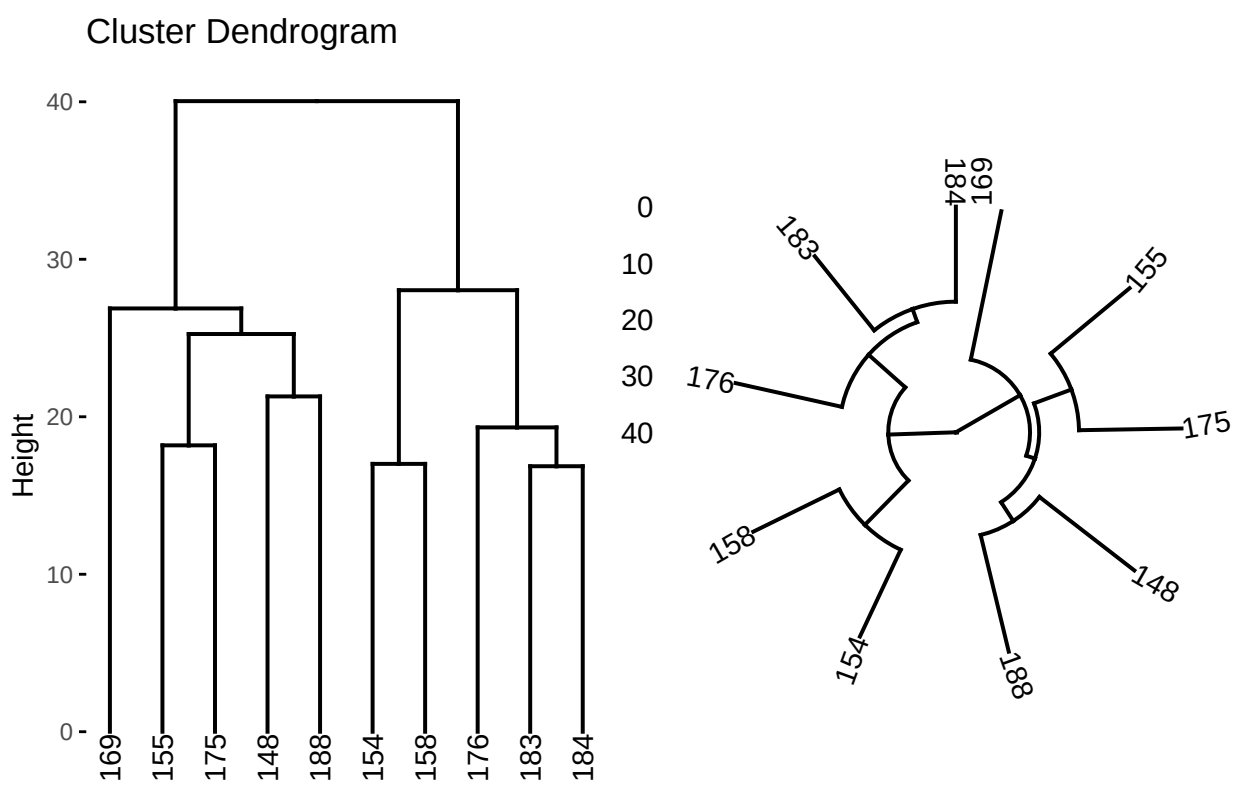
```
p1 <- fviz_dend(dend_cuts$lower[[1]])
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

```
p2 <- fviz_dend(dend_cuts$lower[[1]], type = 'circular')
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

```
gridExtra::grid.arrange(p1, p2, nrow = 1)
```



Modelbased

Helper packages

```r
library(dplyr)
library(ggplot2)

# Modeling packages
library(mclust)
```

```
## Warning: package 'mclust' was built under R version 4.2.2
```

```
## Package 'mclust' version 6.0.0
## Type 'citation("mclust")' for citing this R package in publications.
```

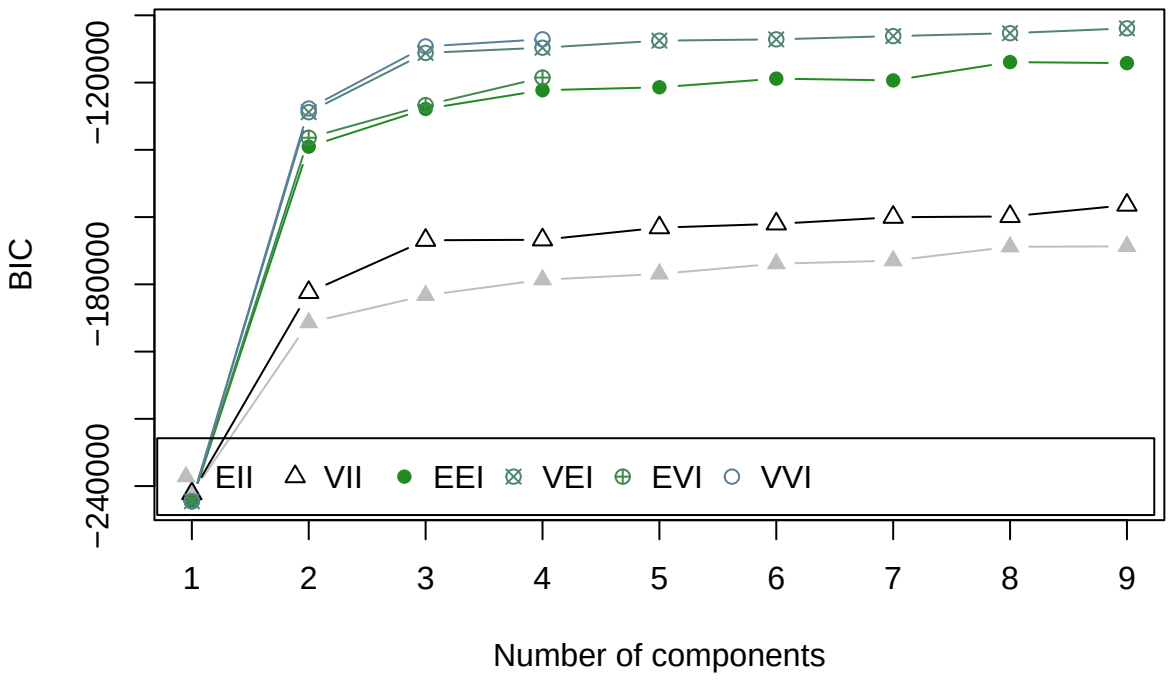
```
##
## Attaching package: 'mclust'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```r
mydata_mc <- Mclust(x_train)

summary(mydata_mc)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEI (diagonal, equal shape) model with 9 components:
##
##  log-likelihood   n   df     BIC     ICL
##       -40533.33 197 4316 -103869 -103869
##
## Clustering table:
##   1   2   3   4   5   6   7   8   9
## 110  20   3   2  12  10  25   9   6
```

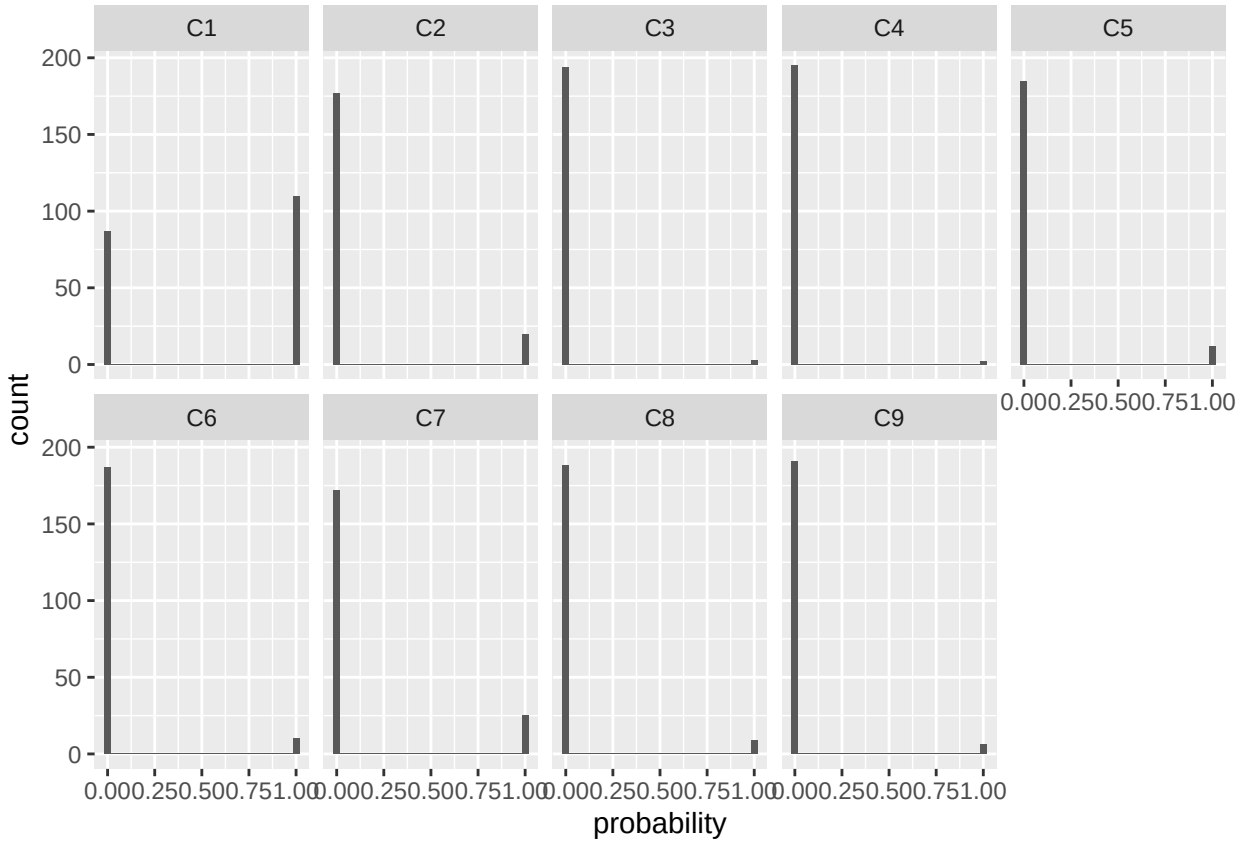
```r
plot(mydata_mc, what = 'BIC',
     legendArgs = list(x = "bottomright", ncol = 9))
```




```
probabilities <- mydata_mc$z
colnames(probabilities) <- paste0('C', 1:9)

probabilities <- probabilities %>%
  as.data.frame() %>%
  mutate(id = row_number()) %>%
  tidyr::gather(cluster, probability, -id)
```

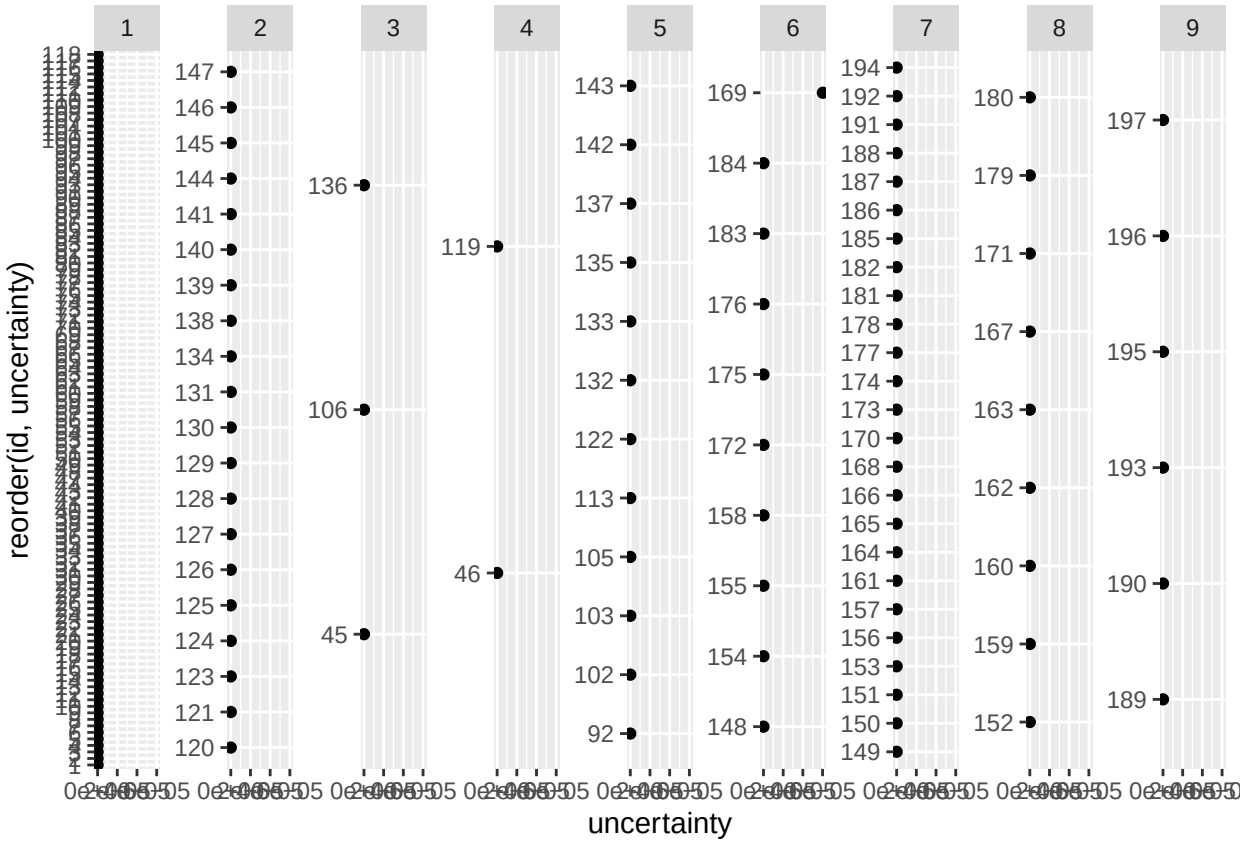
```
ggplot(probabilities, aes(probability)) +
  geom_histogram() +
  facet_wrap(~ cluster, nrow = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



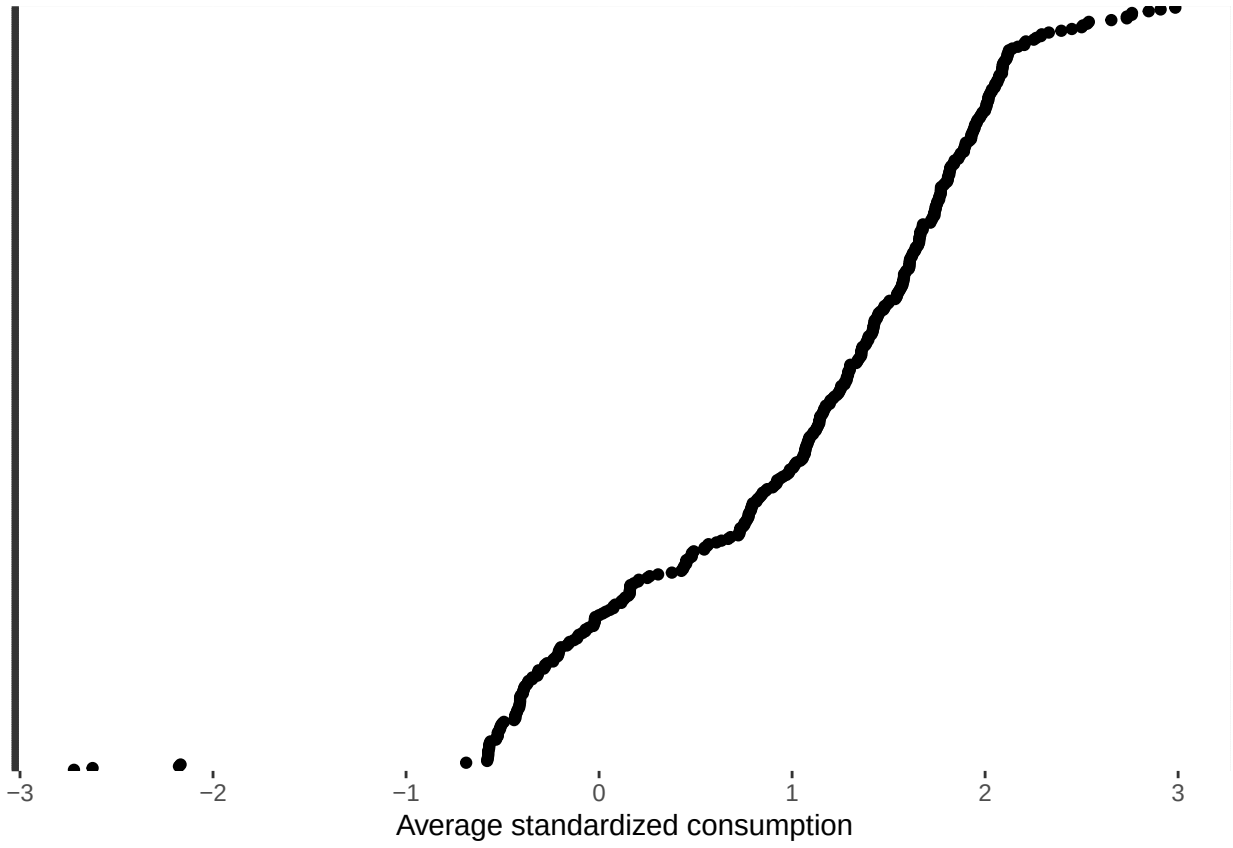

```
uncertainty <- data.frame(
  id = 1:nrow(x_train),
  cluster = mydata_mc$classification,
  uncertainty = mydata_mc$uncertainty
)
```

```
uncertainty %>%
  group_by(cluster) %>%
  filter(uncertainty > -0.25) %>%
  ggplot(aes(uncertainty, reorder(id, uncertainty))) +
  geom_point() +
  facet_wrap(~ cluster, scales = 'free_y', nrow = 1)
```




```
cluster <- x_train %>%
  scale() %>%
  as.data.frame() %>%
  mutate(cluster = mydata_mc$classification) %>%
  filter(cluster == 6) %>%
  select(-cluster)
```

```
cluster %>%
  tidyr::gather(product, std_count) %>%
  group_by(product) %>%
  summarize(avg = mean(std_count)) %>%
  ggplot(aes(avg, reorder(product, avg))) +
  geom_point() +
  labs(x = "Average standardized consumption", y = NULL) +
  theme(axis.text.y=element_blank())
```


−3
−2
−1
0
1
2
3
Average standardized consumption