clusterana.R

jaswinder

2023-04-08

```
rm(list = ls())
library(factoextra)
```

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

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

```
library(cluster)
data = read.csv("fraud_cluster.csv")

rownames(data) = data$state
data = data[-1]
data
```

```
##                     rate2021   rate2020  rate2019  rate2018   rate2017
## Andhra Pradesh    0.66400000 0.60505529 0.6420997 0.6072908 0.57679914
## Arunachal Pradesh 0.17021277 0.86666667 0.0000000 0.2857143 0.00000000
## Assam             0.14527445 0.06855524 0.1089198 0.1923838 0.04285714
## Bihar             0.73106865 0.80555556 0.8038095 0.9385027 0.91685912
## Chhattisgarh      0.21306818 0.25252525 0.2685714 0.1654676 0.27485380
## Goa               0.63888889 0.62500000 0.6000000 0.3793103 0.61538461
## Gujarat           0.63020833 0.68199532 0.4630102 0.5712251 0.66593886
## Haryana           0.12218650 0.23932927 0.2677305 0.3277512 0.08531746
## Himachal Pradesh  0.41428571 0.19387755 0.1973684 0.2608696 0.14285714
## Jharkhand         0.85309549 0.88787375 0.8803653 0.8419355 0.63888889
## Karnataka         0.88938053 0.90121963 0.9468386 0.9318376 0.87082546
## Kerala            0.36261981 0.22535211 0.2182410 0.2735294 0.24062500
## Madhya Pradesh    0.41426146 0.41773963 0.2956811 0.3108108 0.31632653
## Maharashtra       0.57874865 0.62099709 0.7149185 0.5690686 0.60238624
## Manipur           0.23880597 0.50632911 0.0000000 0.4827586 0.47297297
## Meghalaya         0.52336449 0.57042253 0.5617978 0.4729730 0.38461538
## Mizoram           0.00000000 0.23076923 0.0000000 0.0000000 0.70000000
## Nagaland          0.62500000 0.62500000 0.0000000 0.0000000 0.00000000
## Odisha            0.72852234 0.71465562 0.6734007 0.6002372 0.64441748
## Punjab            0.34845735 0.43386243 0.3333333 0.2008368 0.30113636
## Rajasthan         0.50465426 0.47341211 0.5323496 0.4519928 0.25383436
## Sikkim            0.00000000 0.00000000 0.0000000 0.0000000 0.00000000
## Tamil Nadu        0.55947955 0.17135550 0.1896104 0.1864407 0.22368421
## Telangana         0.84344366 0.88296178 0.7480491 0.6074689 0.43755170
## Tripura           0.20833333 0.32352941 0.1000000 0.4000000 0.42857143
## Uttar Pradesh     0.46132065 0.42119492 0.3108795 0.3743631 0.69402535
## Uttarakhand       0.35654596 0.40329218 0.2800000 0.2690058 0.43548387
## West Bengal       0.09161793 0.10112360 0.2029851 0.2029851 0.07922535
```

```r
data.scaled = scale(data)
data.scaled
```

```
##                       rate2021      rate2020   rate2019    rate2018    rate2017
## Andhra Pradesh     0.85707927  0.4919540102  0.9322812  0.85663150  0.67573910
## Arunachal Pradesh -1.03132258  1.4680432806 -1.2619403 -0.40796064 -1.46208125
## Assam             -1.12669474 -1.5097630152 -0.8897331 -0.77498069 -1.30323761
## Bihar              1.11357144  1.2400337006  1.4848857  2.15911445  1.93612091
## Chhattisgarh      -0.86742963 -0.8233587837 -0.3441619 -0.88082783 -0.44337646
## Goa                0.76104627  0.5663690293  0.7884156 -0.03989633  0.81875043
## Gujarat            0.72784903  0.7790223333  0.3202859  0.71480385  1.00612225
## Haryana           -1.21499055 -0.8725938729 -0.3470356 -0.24265148 -1.14586475
## Himachal Pradesh  -0.09790880 -1.0421772226 -0.5874811 -0.50566193 -0.93260246
## Jharkhand          1.58024149  1.5471683190  1.7464967  1.77936623  0.90586553
## Karnataka          1.71900721  1.5969626464  1.9736533  2.13290437  1.76550397
## Kerala            -0.29549591 -0.9247435588 -0.5161539 -0.45587738 -0.57024042
## Madhya Pradesh    -0.09800156 -0.2069330692 -0.2515213 -0.30926915 -0.28966394
## Maharashtra        0.53105059  0.5514339048  1.1811219  0.70632374  0.77057387
## Manipur           -0.76900001  0.1236001377 -1.2619403  0.36691160  0.29092283
## Meghalaya          0.31924365  0.3627369128  0.6578686  0.32842977 -0.03656145
## Mizoram           -1.68227116 -0.9045319347 -1.2619403 -1.53152540  1.13236479
## Nagaland           0.70793068  0.5663690293 -1.2619403 -1.53152540 -1.46208125
## Odisha             1.10383353  0.9008800487  1.0392448  0.82889356  0.92635642
## Punjab            -0.34965772 -0.1467778182 -0.1228537 -0.74173930 -0.34596404
## Rajasthan          0.24768968  0.0007846353  0.5572366  0.24592555 -0.52128191
## Sikkim            -1.68227116 -1.7655471346 -1.2619403 -1.53152540 -1.46208125
## Tamil Nadu         0.45735933 -1.1262084819 -0.6139923 -0.79835179 -0.63302894
## Telangana          1.54332977  1.5288414317  1.2943377  0.85733178  0.15963913
## Tripura           -0.88553722 -0.5584375900 -0.9202143  0.04146526  0.12635510
## Uttar Pradesh      0.08196798 -0.1940411547 -0.1995844 -0.05935142  1.11022062
## Uttarakhand       -0.31872426 -0.2608374455 -0.3051075 -0.47366619  0.15197504
## West Bengal       -1.33189460 -1.3882483377 -0.5682875 -0.73329133 -1.16844425
## attr(,"scaled:center")
##  rate2021  rate2020  rate2019  rate2018  rate2017
## 0.4398873 0.4732018 0.3692843 0.3894557 0.3944799
## attr(,"scaled:scale")
##  rate2021  rate2020  rate2019  rate2018  rate2017
## 0.2614842 0.2680199 0.2926321 0.2542927 0.2698071
```

```r
colMeans(data)
```

```
##  rate2021  rate2020  rate2019  rate2018  rate2017
## 0.4398873 0.4732018 0.3692843 0.3894557 0.3944799
```

```r
cov(data)
```

```
##              rate2021   rate2020   rate2019   rate2018   rate2017
## rate2021 0.06837398 0.05413283 0.06392104 0.04979576 0.04115391
## rate2020 0.05413283 0.07183468 0.05374964 0.05120310 0.04087621
## rate2019 0.06392104 0.05374964 0.08563356 0.06430223 0.05374470
## rate2018 0.04979576 0.05120310 0.06430223 0.06466476 0.04971743
## rate2017 0.04115391 0.04087621 0.05374470 0.04971743 0.07279588
```

```r
sqrt( cov(data) )
```

```
##             rate2021  rate2020  rate2019  rate2018  rate2017
```

```
## rate2021 0.2614842 0.2326646 0.2528261 0.2231496 0.2028643
## rate2020 0.2326646 0.2680199 0.2318397 0.2262810 0.2021787
## rate2019 0.2528261 0.2318397 0.2926321 0.2535788 0.2318290
## rate2018 0.2231496 0.2262810 0.2535788 0.2542927 0.2229741
## rate2017 0.2028643 0.2021787 0.2318290 0.2229741 0.2698071
```

```
cov(data.scaled)
```

```
##           rate2021  rate2020  rate2019  rate2018  rate2017
## rate2021 1.0000000 0.7724105 0.8353652 0.7488814 0.5833272
## rate2020 0.7724105 1.0000000 0.6853090 0.7512687 0.5652625
## rate2019 0.8353652 0.6853090 1.0000000 0.8641123 0.6807070
## rate2018 0.7488814 0.7512687 0.8641123 1.0000000 0.7246385
## rate2017 0.5833272 0.5652625 0.6807070 0.7246385 1.0000000
```

```
cor(data)
```

```
##           rate2021  rate2020  rate2019  rate2018  rate2017
## rate2021 1.0000000 0.7724105 0.8353652 0.7488814 0.5833272
## rate2020 0.7724105 1.0000000 0.6853090 0.7512687 0.5652625
## rate2019 0.8353652 0.6853090 1.0000000 0.8641123 0.6807070
## rate2018 0.7488814 0.7512687 0.8641123 1.0000000 0.7246385
## rate2017 0.5833272 0.5652625 0.6807070 0.7246385 1.0000000
```

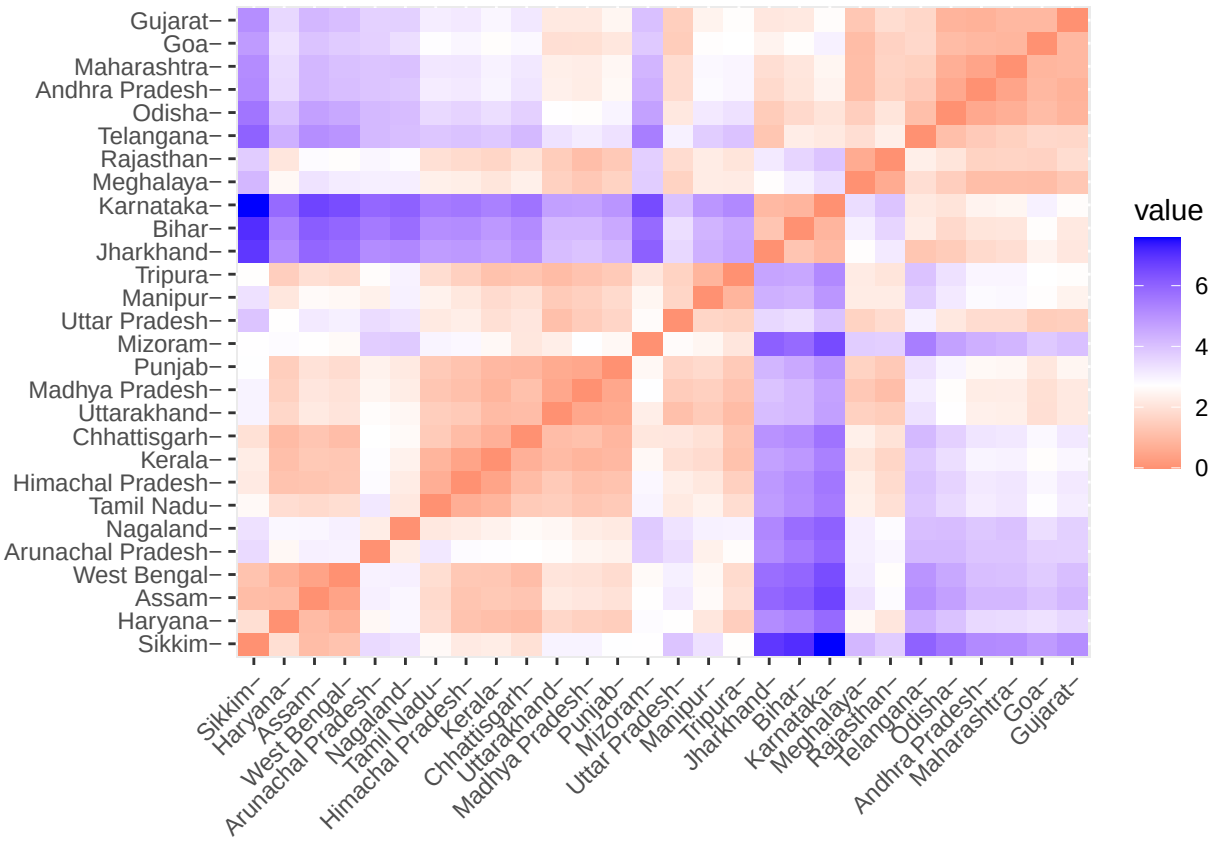
```
cov(data)
```

```
##            rate2021   rate2020   rate2019   rate2018   rate2017
## rate2021 0.06837398 0.05413283 0.06392104 0.04979576 0.04115391
## rate2020 0.05413283 0.07183468 0.05374964 0.05120310 0.04087621
## rate2019 0.06392104 0.05374964 0.08563356 0.06430223 0.05374470
## rate2018 0.04979576 0.05120310 0.06430223 0.06466476 0.04971743
## rate2017 0.04115391 0.04087621 0.05374470 0.04971743 0.07279588
```

```
dist.eucl = dist(data.scaled, method = "euclidean")
dist = as.matrix(dist.eucl)
dist[1:3,1:3]
```

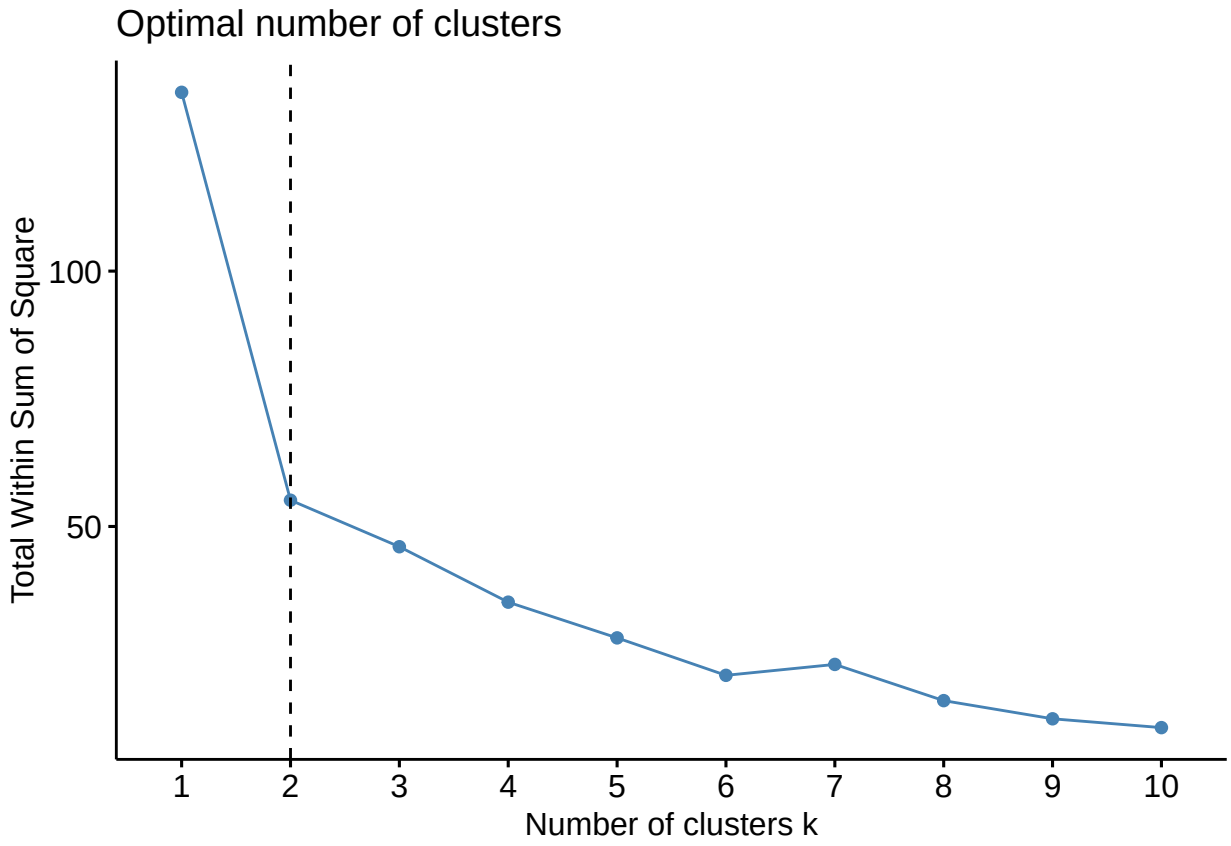
```
##                   Andhra Pradesh Arunachal Pradesh    Assam
## Andhra Pradesh          0.000000          3.937371 4.223798
## Arunachal Pradesh       3.937371          0.000000 3.029010
## Assam                   4.223798          3.029010 0.000000
```

```
fviz_dist(dist.eucl)
```




```
#kemans

fviz_nbclust(data.scaled, kmeans, method = "wss") +
  geom_vline(xintercept = 2, linetype = 2)
```




```
km.res <- kmeans(data.scaled,2 , nstart = 25)
km.res
```

```
## K-means clustering with 2 clusters of sizes 18, 10
##
## Cluster means:
##     rate2021   rate2020   rate2019   rate2018   rate2017
## 1 -0.5697918 -0.5314112 -0.6343662 -0.5735502 -0.4960061
## 2  1.0256252  0.9565402  1.1418591  1.0323903  0.8928110
##
## Clustering vector:
##    Andhra Pradesh Arunachal Pradesh             Assam             Bihar
##                 2                 1                 1                 2
##      Chhattisgarh               Goa           Gujarat           Haryana
##                 1                 2                 2                 1
##  Himachal Pradesh         Jharkhand         Karnataka            Kerala
##                 1                 2                 2                 1
##    Madhya Pradesh       Maharashtra           Manipur         Meghalaya
##                 1                 2                 1                 2
##           Mizoram          Nagaland            Odisha           Punjab
##                 1                 1                 2                 1
##         Rajasthan            Sikkim        Tamil Nadu        Telangana
##                 1                 1                 1                 2
##           Tripura     Uttar Pradesh       Uttarakhand       West Bengal
##                 1                 1                 1                 1
##
## Within cluster sum of squares by cluster:
## [1] 40.44021 14.70286
```

```
##  (between_SS / total_SS =  59.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

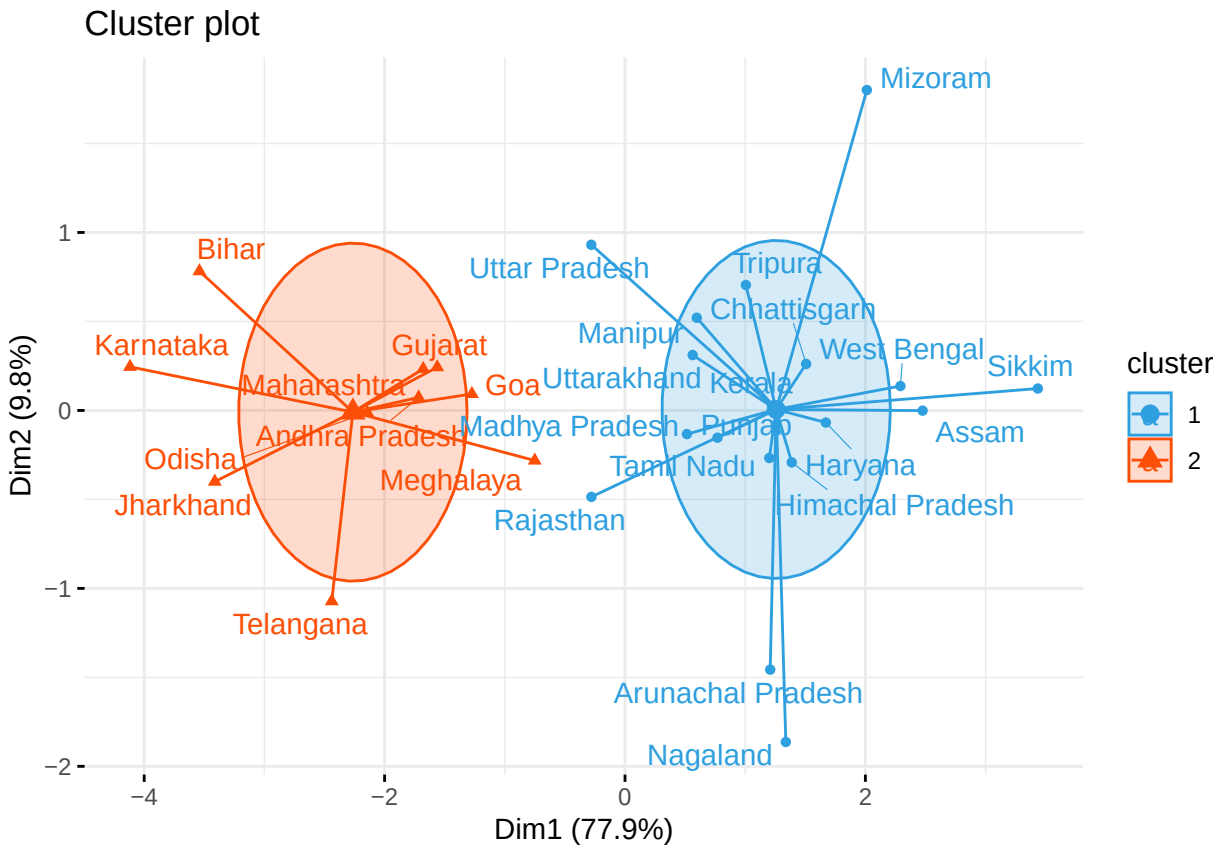
```
aggregate(data, by=list(cluster=km.res$cluster), mean)
```

```
##   cluster  rate2021  rate2020  rate2019  rate2018  rate2017
## 1       1 0.2908958 0.3307730 0.1836483 0.2436061 0.2606539
## 2       2 0.7080721 0.7295737 0.7034289 0.6519850 0.6353667
```

```
dd <- cbind(data, cluster = km.res$cluster)
dd
```

```
##                     rate2021   rate2020  rate2019  rate2018   rate2017 cluster
## Andhra Pradesh    0.66400000 0.60505529 0.6420997 0.6072908 0.57679914       2
## Arunachal Pradesh 0.17021277 0.86666667 0.0000000 0.2857143 0.00000000       1
## Assam             0.14527445 0.06855524 0.1089198 0.1923838 0.04285714       1
## Bihar             0.73106865 0.80555556 0.8038095 0.9385027 0.91685912       2
## Chhattisgarh      0.21306818 0.25252525 0.2685714 0.1654676 0.27485380       1
## Goa               0.63888889 0.62500000 0.6000000 0.3793103 0.61538461       2
## Gujarat           0.63020833 0.68199532 0.4630102 0.5712251 0.66593886       2
## Haryana           0.12218650 0.23932927 0.2677305 0.3277512 0.08531746       1
## Himachal Pradesh  0.41428571 0.19387755 0.1973684 0.2608696 0.14285714       1
## Jharkhand         0.85309549 0.88787375 0.8803653 0.8419355 0.63888889       2
## Karnataka         0.88938053 0.90121963 0.9468386 0.9318376 0.87082546       2
## Kerala            0.36261981 0.22535211 0.2182410 0.2735294 0.24062500       1
## Madhya Pradesh    0.41426146 0.41773963 0.2956811 0.3108108 0.31632653       1
## Maharashtra       0.57874865 0.62099709 0.7149185 0.5690686 0.60238624       2
## Manipur           0.23880597 0.50632911 0.0000000 0.4827586 0.47297297       1
## Meghalaya         0.52336449 0.57042253 0.5617978 0.4729730 0.38461538       2
## Mizoram           0.00000000 0.23076923 0.0000000 0.0000000 0.70000000       1
## Nagaland          0.62500000 0.62500000 0.0000000 0.0000000 0.00000000       1
## Odisha            0.72852234 0.71465562 0.6734007 0.6002372 0.64441748       2
## Punjab            0.34845735 0.43386243 0.3333333 0.2008368 0.30113636       1
## Rajasthan         0.50465426 0.47341211 0.5323496 0.4519928 0.25383436       1
## Sikkim            0.00000000 0.00000000 0.0000000 0.0000000 0.00000000       1
## Tamil Nadu        0.55947955 0.17135550 0.1896104 0.1864407 0.22368421       1
## Telangana         0.84344366 0.88296178 0.7480491 0.6074689 0.43755170       2
## Tripura           0.20833333 0.32352941 0.1000000 0.4000000 0.42857143       1
## Uttar Pradesh     0.46132065 0.42119492 0.3108795 0.3743631 0.69402535       1
## Uttarakhand       0.35654596 0.40329218 0.2800000 0.2690058 0.43548387       1
## West Bengal       0.09161793 0.10112360 0.2029851 0.2029851 0.07922535       1
```

```
fviz_cluster(km.res, data = data,
             palette = c("#2E9FDF", "#FC4E07"),
             ellipse.type = "euclid", # Concentration ellipse
             star.plot = TRUE, # Add segments from centroids to items
             repel = TRUE, # Avoid label overplotting (slow)
             ggtheme = theme_minimal() )
```

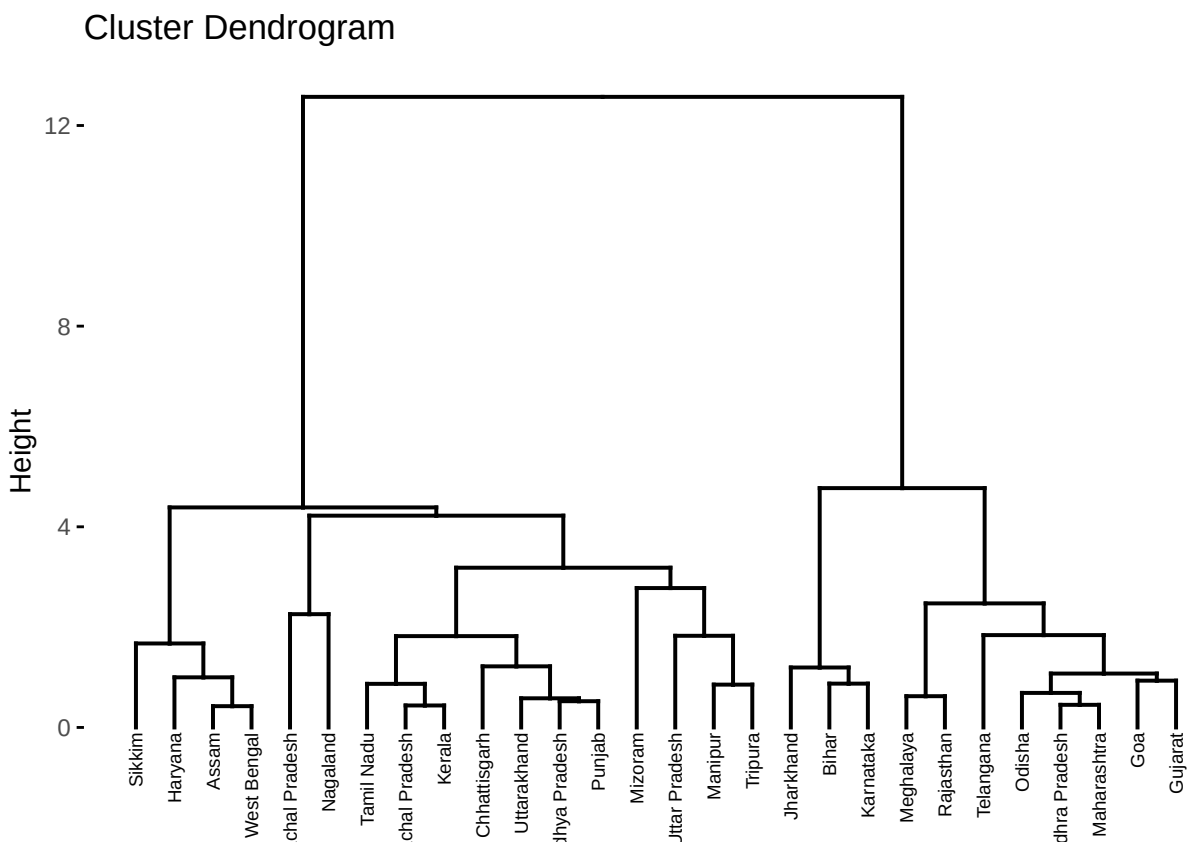



```
## Algglomerative
res.hc <- hclust(d = dist.eucl, method = "ward.D2")
res.hc
```

```
##
## Call:
## hclust(d = dist.eucl, method = "ward.D2")
##
## Cluster method   : ward.D2
## Distance         : euclidean
## Number of objects: 28
```

```
fviz_dend(res.hc, cex = 0.5)
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <https://github.com/kassambara/factoextra/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



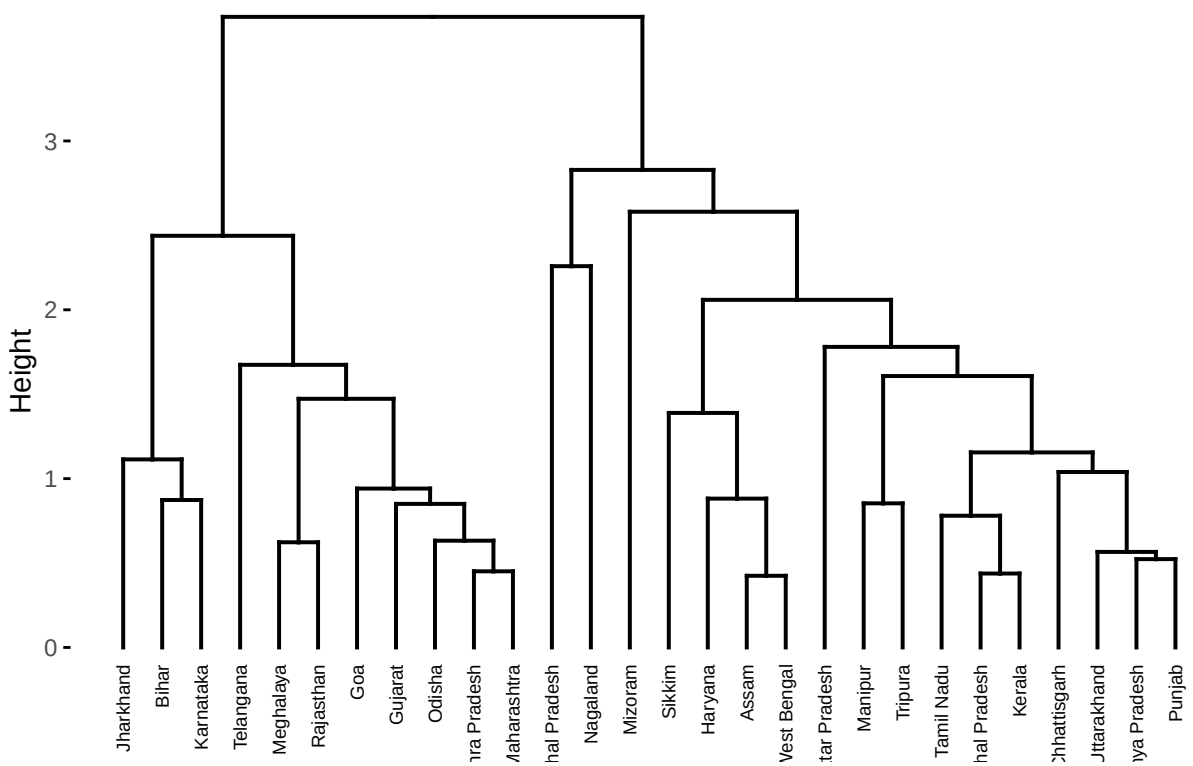

```
res.coph <- cophenetic(res.hc)
cor(dist.eucl, res.coph)
```

```
## [1] 0.6792647
```

```
res.hc2 <- hclust(d = dist.eucl, method = "average")
res.hc2
```

```
##
## Call:
## hclust(d = dist.eucl, method = "average")
##
## Cluster method   : average
## Distance         : euclidean
## Number of objects: 28
```

```
fviz_dend(res.hc2, cex = 0.5)
```




```
res.coph2 <- cophenetic(res.hc2)
cor(dist.eucl, res.coph2)
```

```
## [1] 0.7154447
```

```
# better dendrogram

grp <- cutree(res.hc, k = 2)
head(grp, n = 5)
```

```
##    Andhra Pradesh Arunachal Pradesh             Assam             Bihar
##                 1                 2                 2                 1
##      Chhattisgarh
##                 2
```

```
table(grp)
```

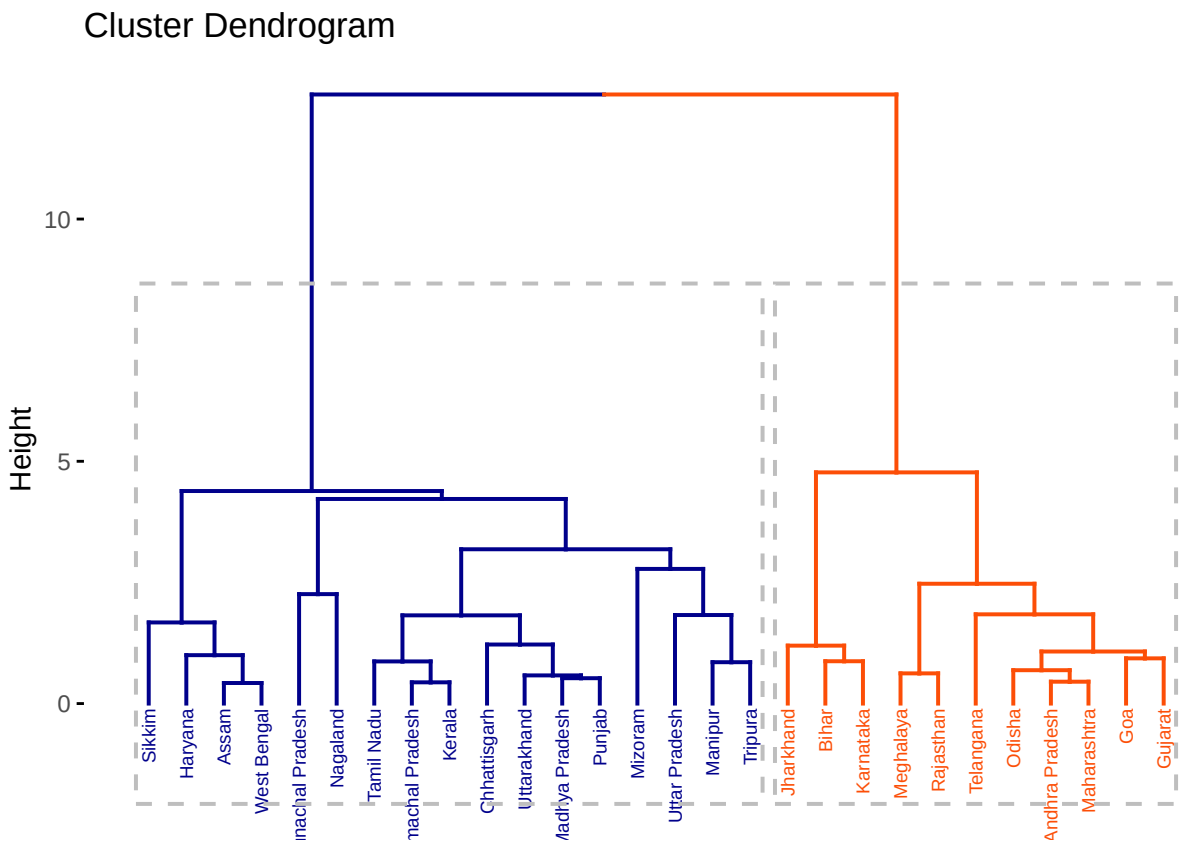
```
## grp
##  1  2
## 11 17
```

```
rownames(data)[grp == 1]
```

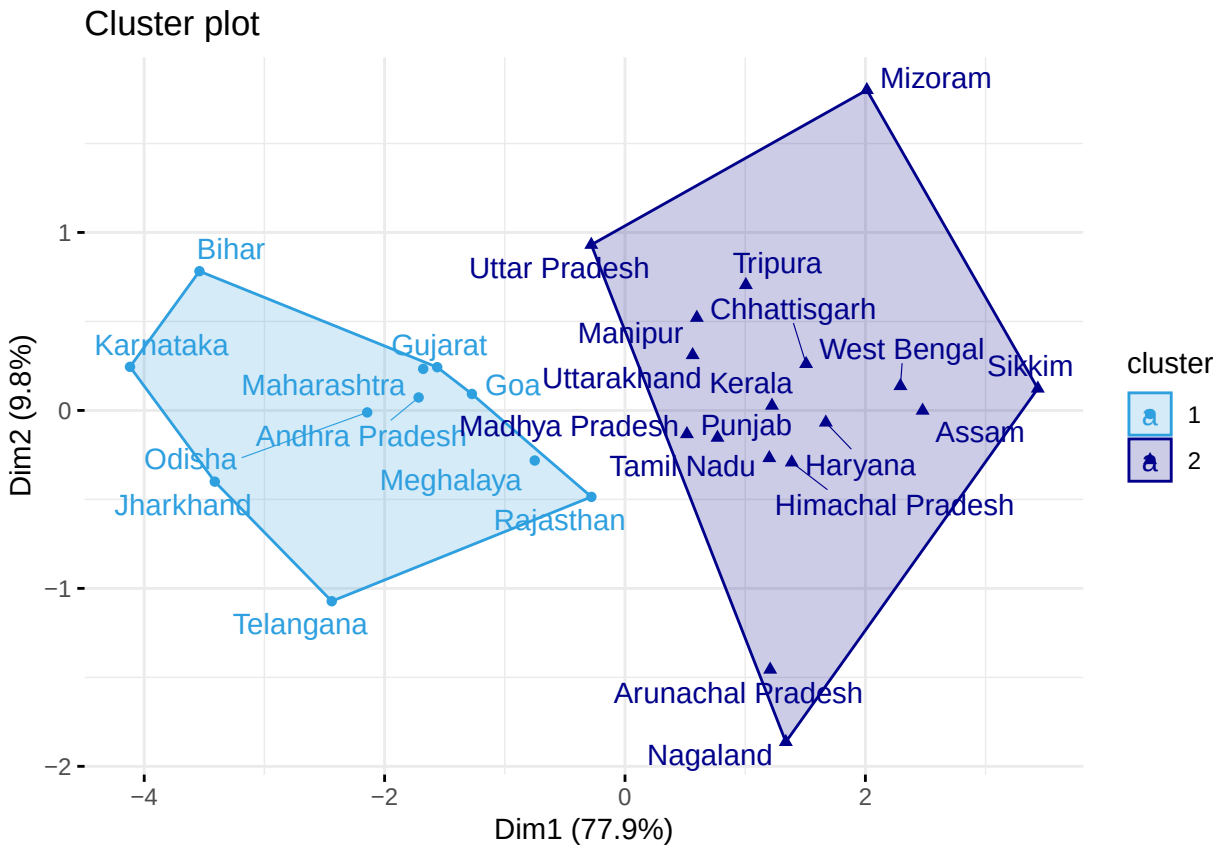
```
##  [1] "Andhra Pradesh" "Bihar"          "Goa"            "Gujarat"
##  [5] "Jharkhand"      "Karnataka"      "Maharashtra"    "Meghalaya"
##  [9] "Odisha"         "Rajasthan"      "Telangana"
```

```
plt = fviz_dend(res.hc, k = 2, # Cut in four groups
          cex = 0.5, # label size
          k_colors = c( "blue4", "#FC4E07"),
```

```
          color_labels_by_k = TRUE, # color labels by groups
          rect = TRUE # Add rectangle around groups
)
show(plt)
```

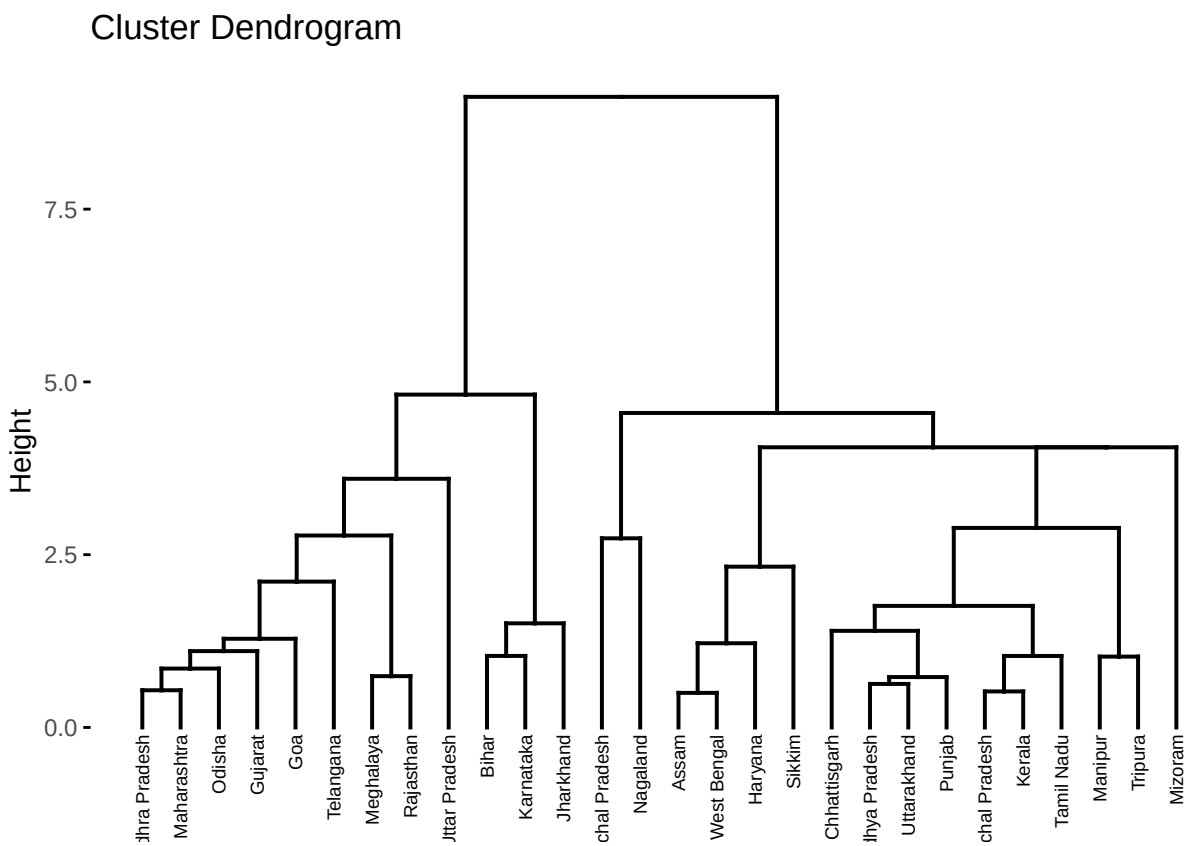



```
fviz_cluster(list(data = data, cluster = grp),
             palette = c("#2E9FDF", "blue4", "#FC4E07"),
             ellipse.type = "convex", # Concentration ellipse
             repel = TRUE, # Avoid label overplotting (slow)
             show.clust.cent = FALSE, ggtheme = theme_minimal())
```

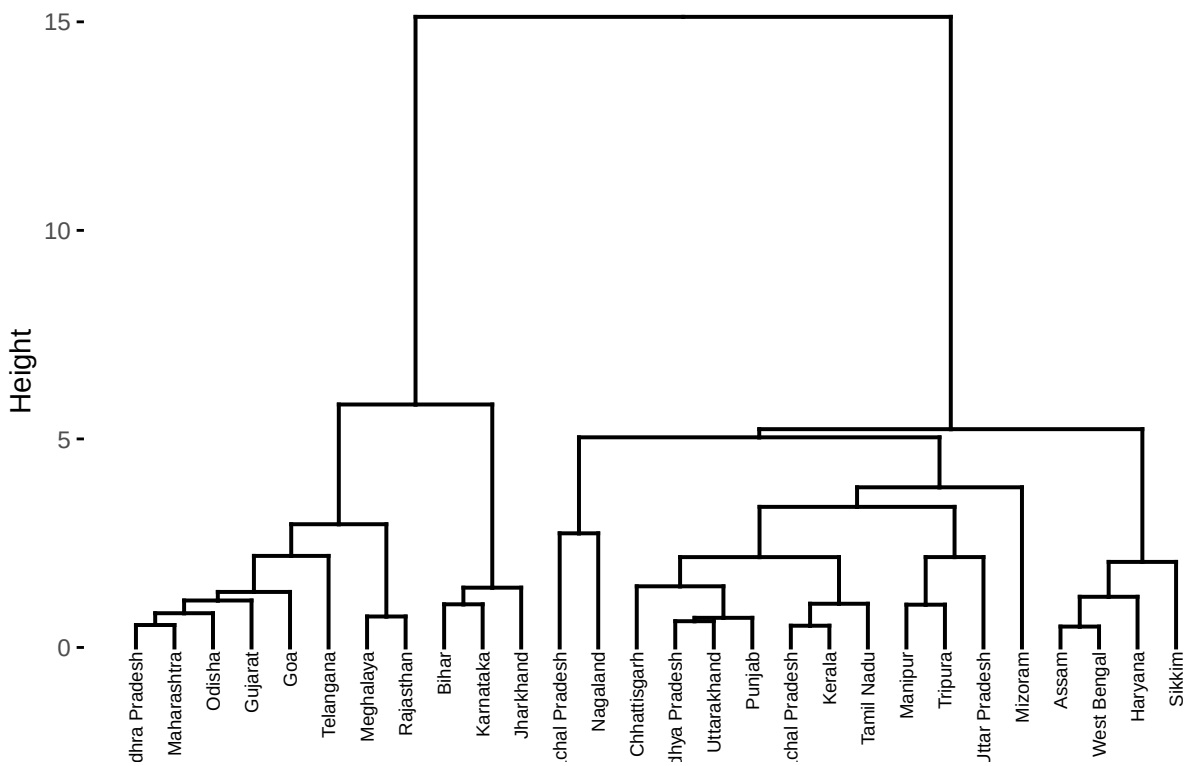



```
#divisive clustering
res.diana <- diana(x = data, # data matrix
                   stand = TRUE, # standardize the data
                   metric = "euclidean")

fviz_dend(res.diana, cex = 0.5)
```




```r
#agglomerative
res.agnes <- agnes(x = data, # data matrix
                   stand = TRUE, # Standardize the data
                   metric = "euclidean", # metric for distance matrix
                   method = "ward" # Linkage method
)
fviz_dend(res.agnes, cex = 0.5)
```




```
res.agnes
```

```
## Call:     agnes(x = data, metric = "euclidean", stand = TRUE, method = "ward")
## Agglomerative coefficient:  0.9176481
## Order of objects:
##  [1] Andhra Pradesh    Maharashtra       Odisha            Gujarat
##  [5] Goa               Telangana         Meghalaya         Rajasthan
##  [9] Bihar             Karnataka         Jharkhand         Arunachal Pradesh
## [13] Nagaland          Chhattisgarh      Madhya Pradesh    Uttarakhand
## [17] Punjab            Himachal Pradesh  Kerala            Tamil Nadu
## [21] Manipur           Tripura           Uttar Pradesh     Mizoram
## [25] Assam             West Bengal       Haryana           Sikkim
## Height (summary):
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5005  0.9233  1.4335  2.4757  2.8459 15.1194
##
## Available components:
## [1] "order"     "height"    "ac"        "merge"     "diss"      "call"
## [7] "method"    "order.lab" "data"
```