

Clustering Analyses: Partitioning

Bryn Bandt-law

2/9/2021

The 'knit' chunk will allow us to 'knit' the Rmarkdown to a pdf for submission

Clustering Technique: Partitioning

read in Group 4 merged data set from out github folder

```
library(readr)

link='https://raw.githubusercontent.com/Public-Policy-COVID/students_merge/main/Merged_data.csv'

data = read.csv(link)
```

reset indexes to R format:

```
row.names(data)=NULL
```

Partitioning

-We will request a particular number of clusters to the algorithm. The algorithm will put every case in one of those clusters and outliers will affect output

For clustering, the variables need to be numeric. Change non-numeric variables (Deaths_COVID & Deaths_total) with integers to numeric.

```
data$Deaths_COVID<-as.numeric(data$Deaths_COVID)

data$Deaths_total<-as.numeric(data$Deaths_total)
```

a. select variables to use for clustering

```
dfClus=data[,c('Number_of_beds','mask_score','Deaths_COVID','Deaths_total','Number_of_hospitals', "black_total_pct", "white_total_pct")]

summary(dfClus)
```

```
## Number_of_beds      mask_score      Deaths_COVID      Deaths_total
## Min.   :  0.0      Min.   :2.470      Min.   :  0      Min.   :  0
## 1st Qu.: 25.0      1st Qu.:3.301      1st Qu.:  0      1st Qu.:  0
## Median : 131.0      Median :3.464      Median : 22      Median : 637
## Mean   : 885.4      Mean   :3.428      Mean   : 206      Mean   : 2896
## 3rd Qu.: 553.0      3rd Qu.:3.591      3rd Qu.: 128      3rd Qu.: 2537
## Max.   :26672.0      Max.   :3.822      Max.   :8034      Max.   :75463
## Number_of_hospitals black_total_pct white_total_pct
## Min.   :  0          Min.   : 0.000      Min.   :49.28
## 1st Qu.:  1          1st Qu.: 0.770      1st Qu.:82.16
## Median :  2          Median : 1.260      Median :88.64
## Mean   :  5          Mean   : 2.318      Mean   :85.50
## 3rd Qu.:  4          3rd Qu.: 2.620      3rd Qu.:91.84
```

```
## Max.      :112          Max.      :14.770    Max.      :96.13
```

b. rescale units

```
dfClus=scale(dfClus)
summary(dfClus)
```

```
## Number_of_beds      mask_score      Deaths_COVID      Deaths_total
## Min.      :-0.3334    Min.      :-4.2726    Min.      :-0.2704    Min.      :-0.37704
## 1st Qu.: -0.3240    1st Qu.: -0.5659    1st Qu.: -0.2704    1st Qu.: -0.37704
## Median : -0.2841    Median :  0.1612    Median : -0.2415    Median : -0.29411
## Mean   :  0.0000    Mean   :  0.0000    Mean   :  0.0000    Mean   :  0.00000
## 3rd Qu.: -0.1252    3rd Qu.:  0.7277    3rd Qu.: -0.1024    3rd Qu.: -0.04674
## Max.    :  9.7118    Max.    :  1.7581    Max.    :10.2736    Max.    :  9.44771

## Number_of_hospitals black_total_pct  white_total_pct
## Min.      :-0.44686    Min.      :-0.8976    Min.      :-3.8920
## 1st Qu.: -0.35749    1st Qu.: -0.5994    1st Qu.: -0.3585
## Median : -0.26812    Median : -0.4097    Median :  0.3379
## Mean   :  0.00000    Mean   :  0.0000    Mean   :  0.0000
## 3rd Qu.: -0.08937    3rd Qu.:  0.1169    3rd Qu.:  0.6818
## Max.    :  9.56284    Max.    :  4.8214    Max.    :  1.1428
```

c. rename subset indexes and verify input:

```
row.names(dfClus)=data$Location
head(dfClus)
```

```
##           Number_of_beds mask_score Deaths_COVID Deaths_total
## Alameda_CA      1.0476322  1.0666781    0.4816583    1.04310240
## Alpine_CA      -0.3334445 -0.6640201   -0.2703586   -0.37704284
## Amador_CA      -0.3138601 -0.1465949   -0.2296736   -0.32301275
## Butte_CA       -0.1251719 -0.2090427   -0.1378042   -0.07590643
## Calaveras_CA   -0.3240289 -0.6104934   -0.2546096   -0.32691854
## Colusa_CA      -0.3153666  0.1790262   -0.2546096   -0.36194046
##           Number_of_hospitals black_total_pct white_total_pct
## Alameda_CA      1.51932965    3.3732540    -3.89196747
## Alpine_CA      -0.44686166   -0.7620594   -1.88666366
## Amador_CA      -0.35748933    0.1401204    0.44640953
## Butte_CA       0.08937233   -0.1618969    0.01654805
## Calaveras_CA   -0.35748933   -0.4794022    0.58611451
## Colusa_CA      -0.35748933   -0.3903458    0.60223432
```

d. set random seed for replicability of results

```
set.seed(999)
```

e. designate distance method and compute distance matrix

```
library(cluster)
dfClus_D=cluster::daisy(x=dfClus)
```

f. For the partitioning technique, we need to indicate the number of clusters required

```
NumCluster=4
res.pam = pam(x=dfClus_D,
              k = NumCluster,
              cluster.only = F)
```

g. Append the clustering results to the dataframe (data)

```
data$pam=as.factor(res.pam$clustering)
```

h. query the data frame (we will create a table to see the number of counties per cluster and will look at King, County, WA)

```
table(data$pam)
```

```
##  
##  1  2  3  4  
## 14 50 39 30
```

```
data[data$Location=="King_WA",'pam']
```

```
## [1] 1  
## Levels: 1 2 3 4
```

Evaluate results

(a)create average silhouettes

```
library(factoextra)
```

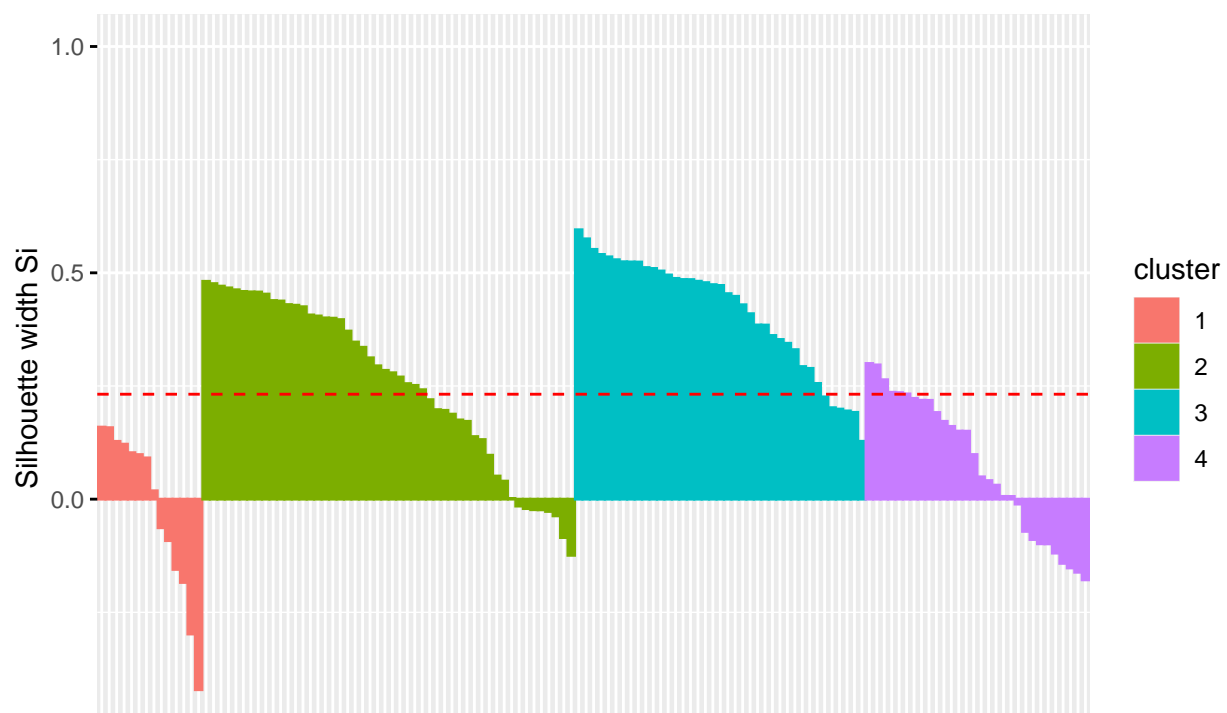
```
## Loading required package: ggplot2
```

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

```
fviz_silhouette(res.pam)
```

```
##   cluster size ave.sil.width  
## 1         1  14        -0.02  
## 2         2  50         0.26  
## 3         3  39         0.41  
## 4         4  30         0.07
```

Clusters silhouette plot
Average silhouette width: 0.23



the average silhouette width is .23

(b) detect anomalies: -save individual silhouettes

```
pamEval=data.frame(res.pam$silinfo$widths)
head(pamEval)
```

```
##           cluster neighbor sil_width
## Alameda_CA      1         4 0.1591567
## San Bernardino_CA 1         4 0.1579688
## Los Angeles_CA   1         4 0.1277071
## San Diego_CA     1         4 0.1215627
## Sacramento_CA    1         4 0.1027119
## King_WA          1         4 0.0985324
```

-request negative silhouettes.

A negative silhouettes indicates that the item is poorly clustered

```
pamEval[pamEval$sil_width<0,]
```

```
##           cluster neighbor  sil_width
## Solano_CA      1         4 -0.06312941
## San Francisco_CA 1         4 -0.09158907
## Santa Clara_CA  1         4 -0.15525012
## Contra Costa_CA 1         4 -0.18405269
## San Joaquin_CA  1         4 -0.29794661
## Fresno_CA       1         4 -0.42095642
## Wheeler_OR      2         3 -0.01532344
## Grant_WA        2         3 -0.02085607
## Alpine_CA       2         4 -0.02335024
## Grant_OR        2         3 -0.02363649
## Cowlitz_WA      2         3 -0.02718744
## Franklin_WA     2         3 -0.03694608
## Kittitas_WA     2         3 -0.08484523
## Josephine_OR    2         3 -0.12406058
## Inyo_CA         4         3 -0.01091712
## Santa Barbara_CA 4         3 -0.07124933
## Sonoma_CA       4         3 -0.08919528
## Humboldt_CA     4         3 -0.09860057
## Imperial_CA     4         3 -0.09878916
## Whitman_WA      4         3 -0.11925089
## Lake_CA         4         3 -0.14186724
## Placer_CA       4         3 -0.15197878
## Island_WA       4         3 -0.16148989
## Clark_WA        4         3 -0.17829522
```