HDBSCAN

With Log-Transformed Data

PMS

29 May, 2023

Introduction

The following analysis is an implementation of the HDBSCAN algorithm on individual, log-scaled proximity measures (no auxiliary variables included).

Please note that for the following plots, HDBSCAN considers cluster 1 to be “noise points” and are thus not part of an actual cluster.

Assumptions of the Alogrithm

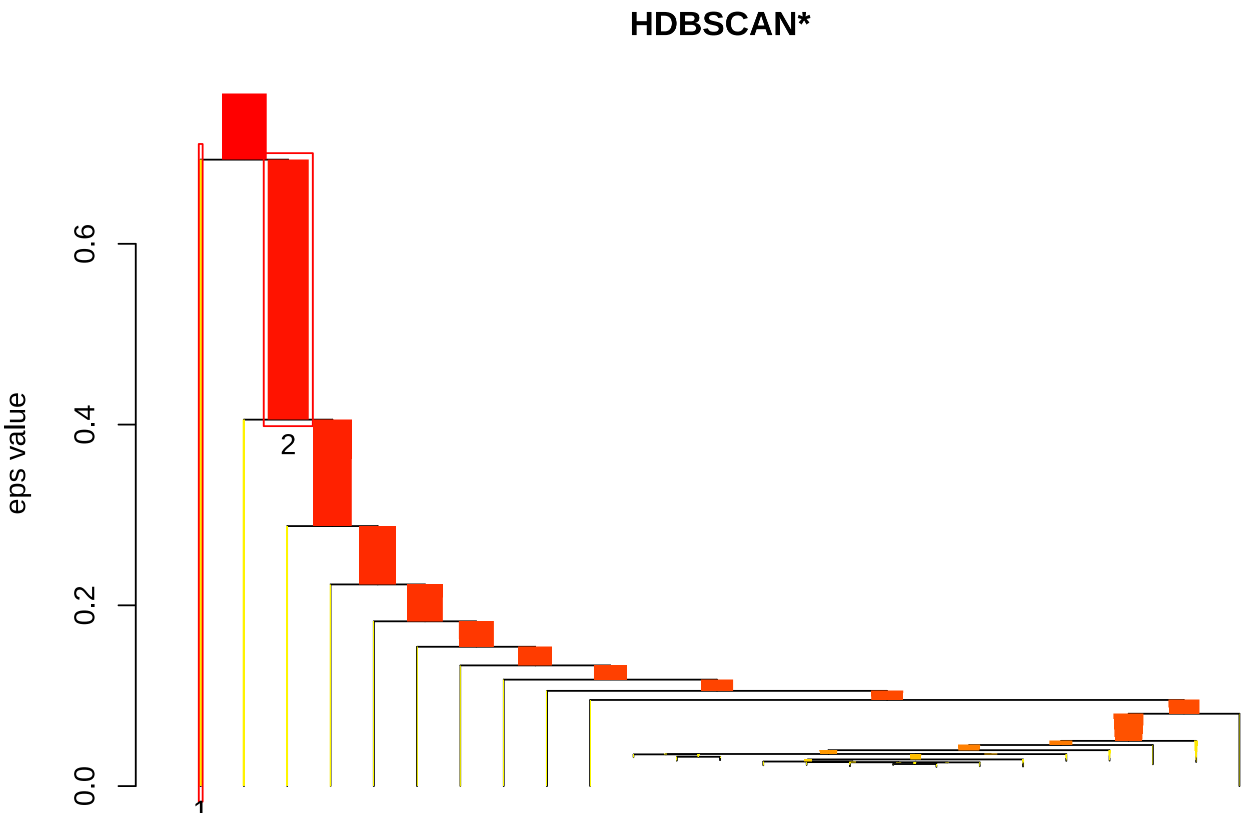
This fast implementation of HDBSCAN (Campello et al., 2013) computes the hierarchical cluster tree representing density estimates along with the stability-based flat cluster extraction. HDBSCAN essentially computes the hierarchy of all DBSCAN* clusterings, and then uses a stability-based extraction method to find optimal cuts in the hierarchy, thus producing a flat solution.

HDBSCAN performs the following steps:

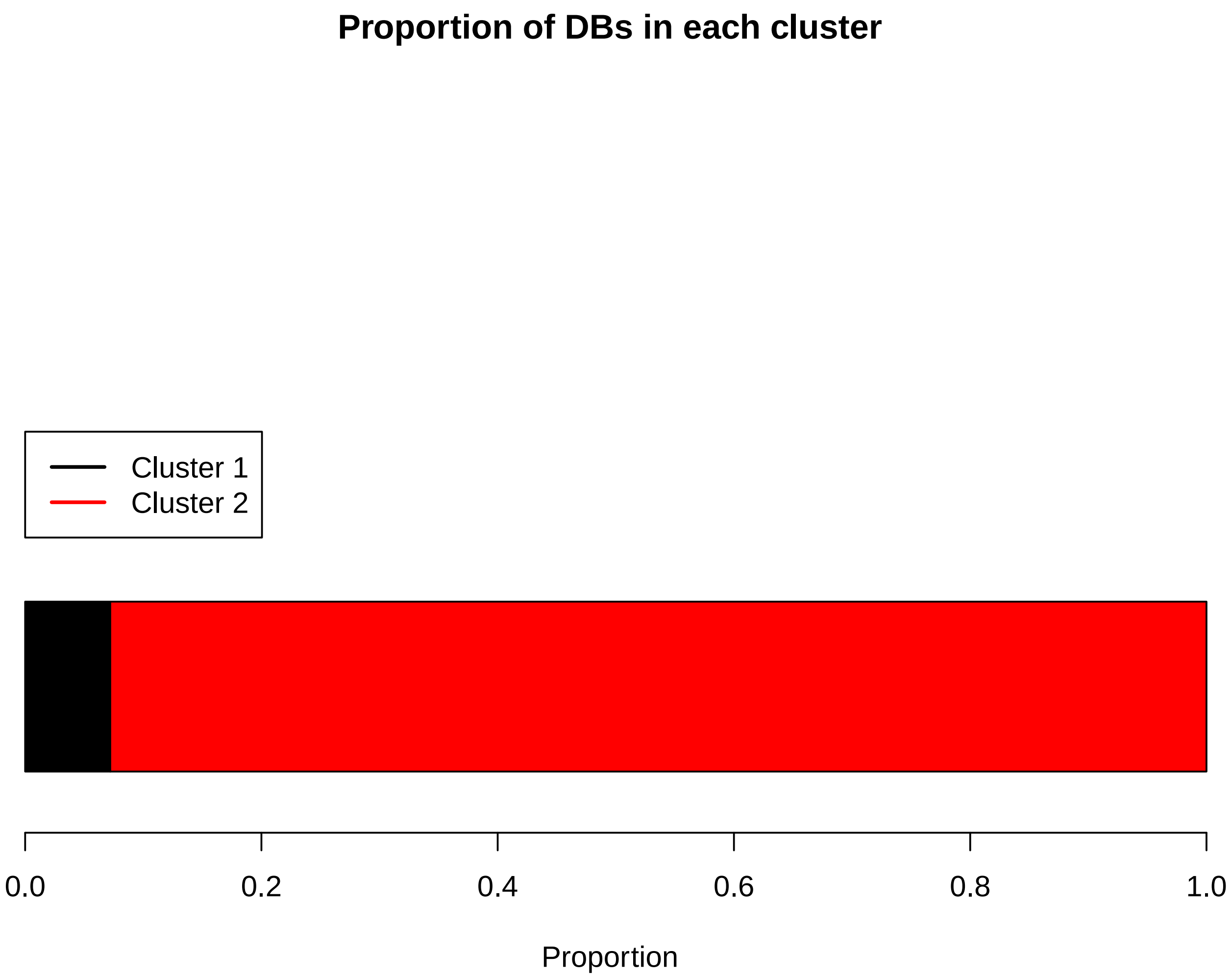
- Compute mutual reachability distance mrd between points (based on distances and core distances).
- Use mdr as a distance measure to construct a minimum spanning tree.
- Prune the tree using stability.
- Extract the clusters.

Amenities

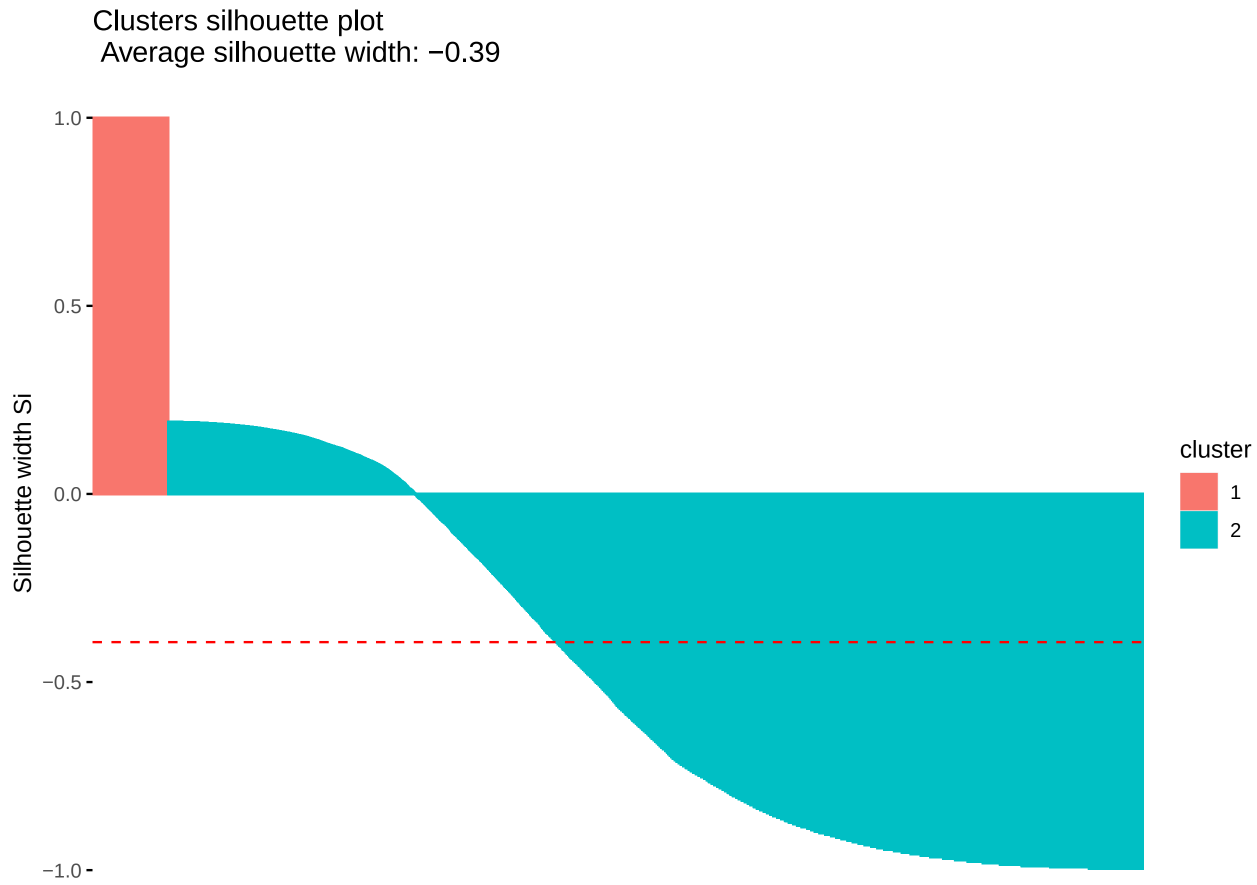
Employment



```
## [1] "Silhouette coefficient: 0.689974748481503"
## [1] "Xie Beni coefficient: 5098.73176887421"
## [1] "Davies Bouldin coefficient: 0.397646393855886"
## [1] "Dunn Index coefficient: 0.00338328122863714"
## [1] "Calinski-Harabasz coefficient: 3655.75180109406"
```



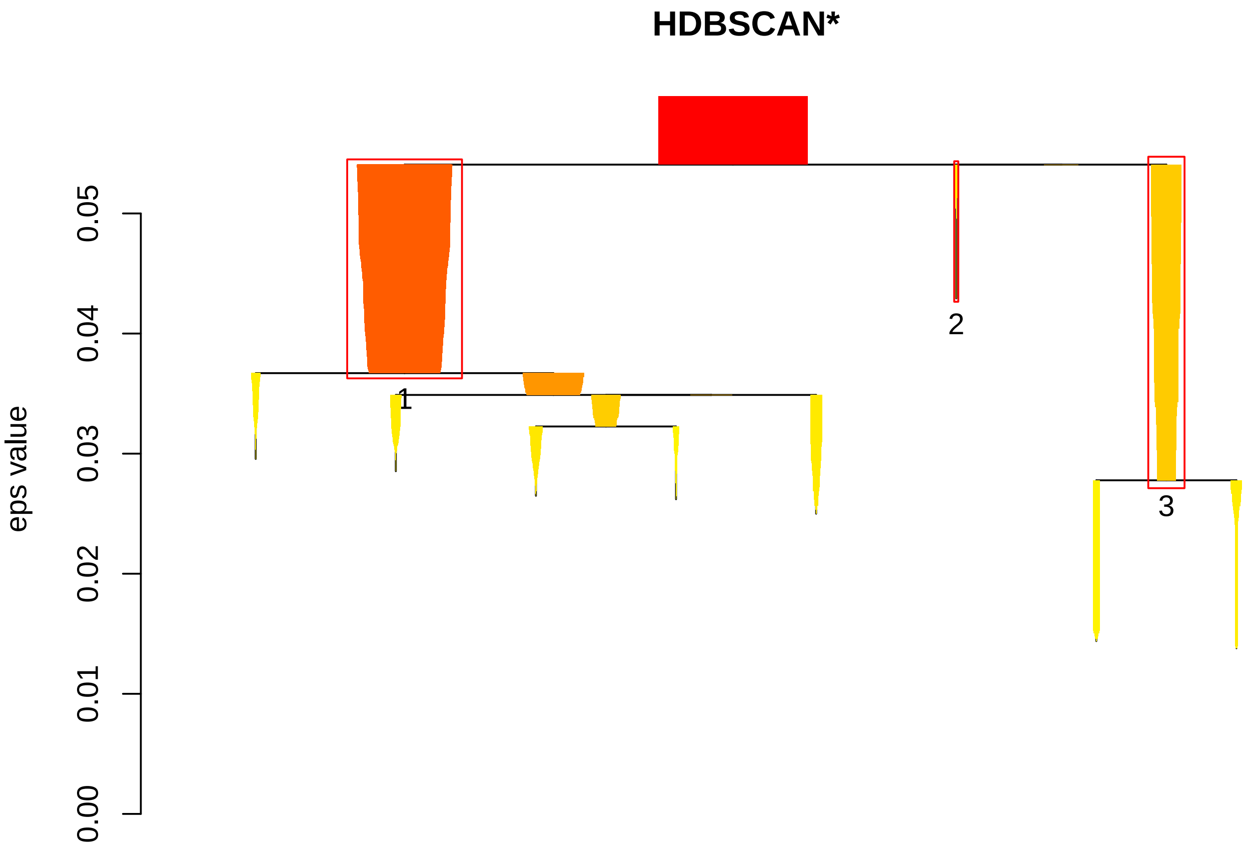
```
## [1] "Segment cutoff values:"
## [1] 0.22985
##   cluster  size ave.sil.width
## 1       1  1521           1.0
## 2       2 19598          -0.5
```



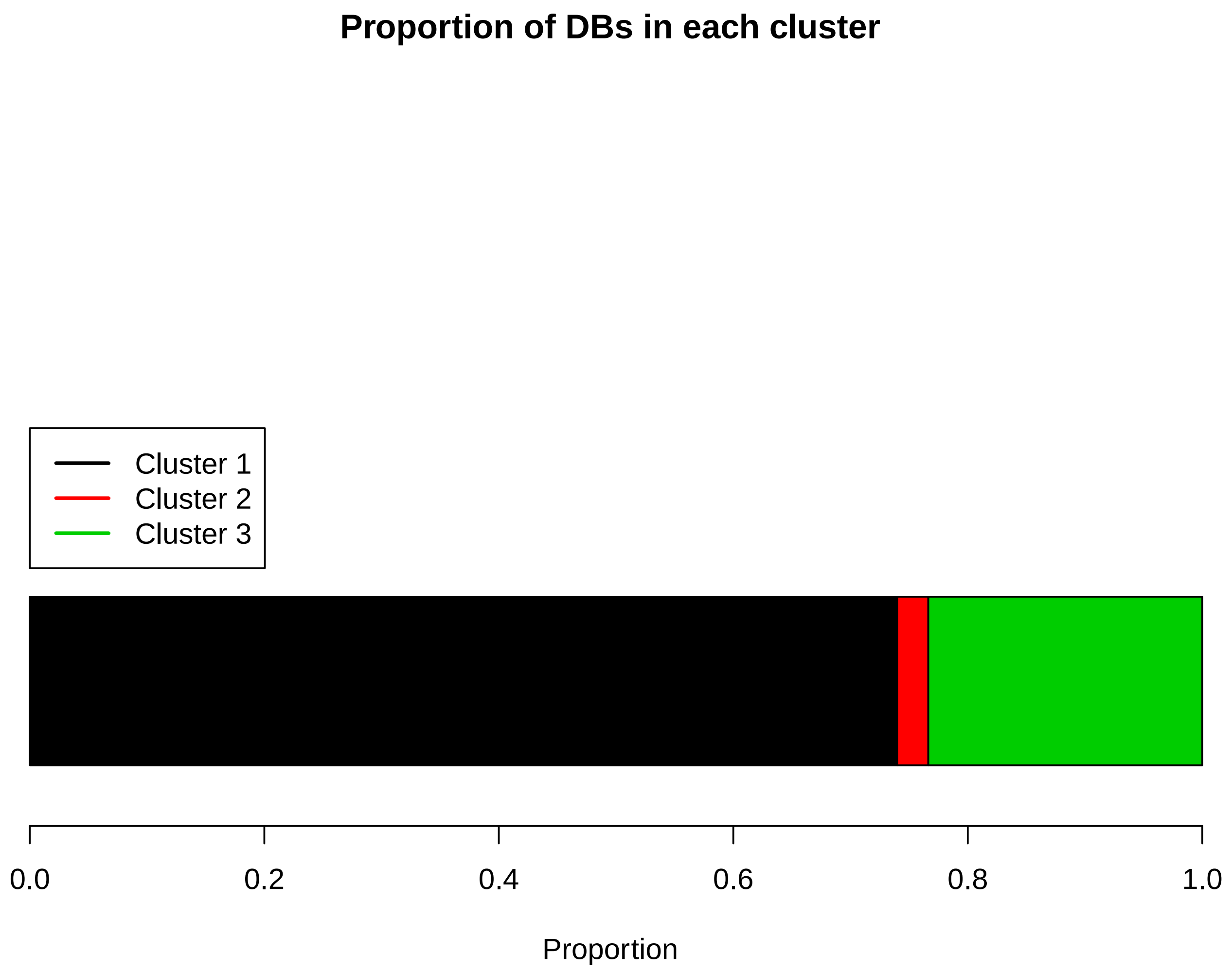
```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##       1777     22706
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       72.7      71.4
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   242835.3  240171.7
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2
##         763      9864
## B       749      9563
## D       198      2462
## K        67       817
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2
##      0.226     0.226
##
##
##
##  Provinces:
##                      Cluster 1 Cluster 2
## Alberta                     64       699
## BritishColumbia             73      1126
## NewBrunswick                16       191
## NorthwestTerritories         0        16
## NovaScotia                  68       711
## Ontario                    296      3562
## Quebec                     111      1267
## Saskatchewan                 6       115
## NA's                      1143     15019
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0      1612     20506
## 1       122      1693
## 2        23       244
## F        20       263
##
##
##
##  PMS_prox_idx_emp :
##  Cluster 1 Cluster 2
##    0.02543   0.02566
```

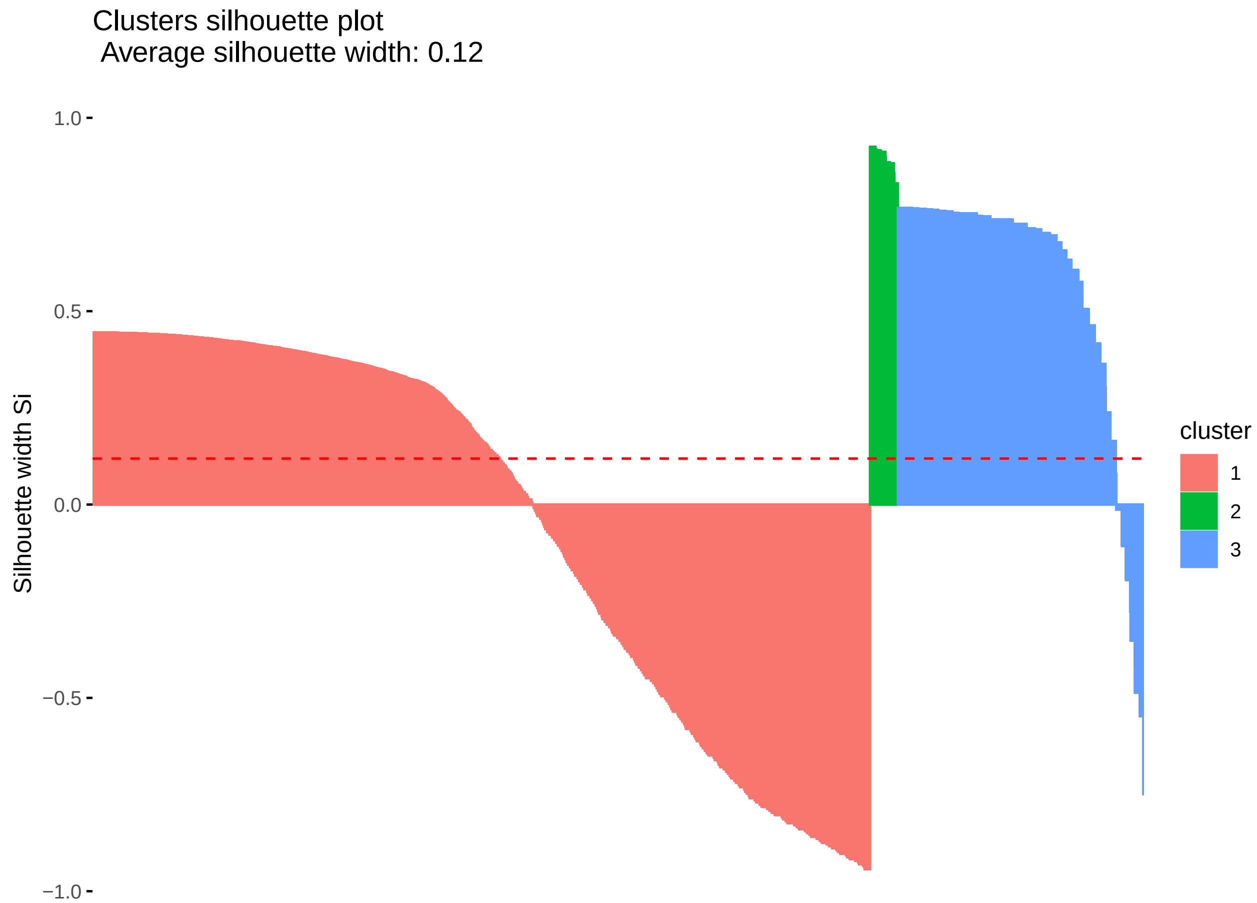
Pharmacy



```
## [1] "Silhouette coefficient: 0.440268773914686"
## [1] "Xie Beni coefficient: Inf"
## [1] "Davies Bouldin coefficient: 0.796395288967002"
## [1] "Dunn Index coefficient: 0"
## [1] "Calinski-Harabasz coefficient: 4571.01854374982"
```



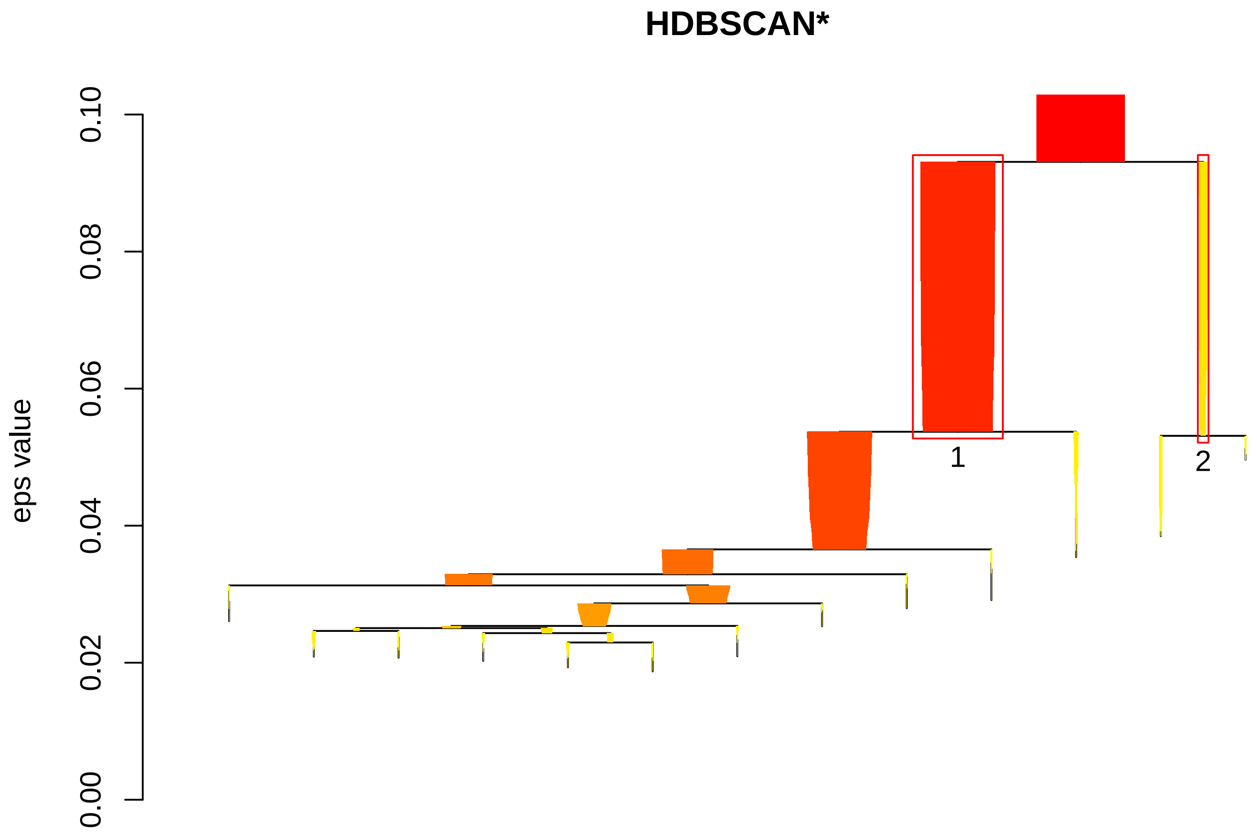
```
## [1] "Segment cutoff values:"
## [1] 0.01175
## [1] 0.0525
##   cluster size ave.sil.width
## 1       1 5676         -0.05
## 2       2  204          0.89
## 3       3 1793          0.57
```



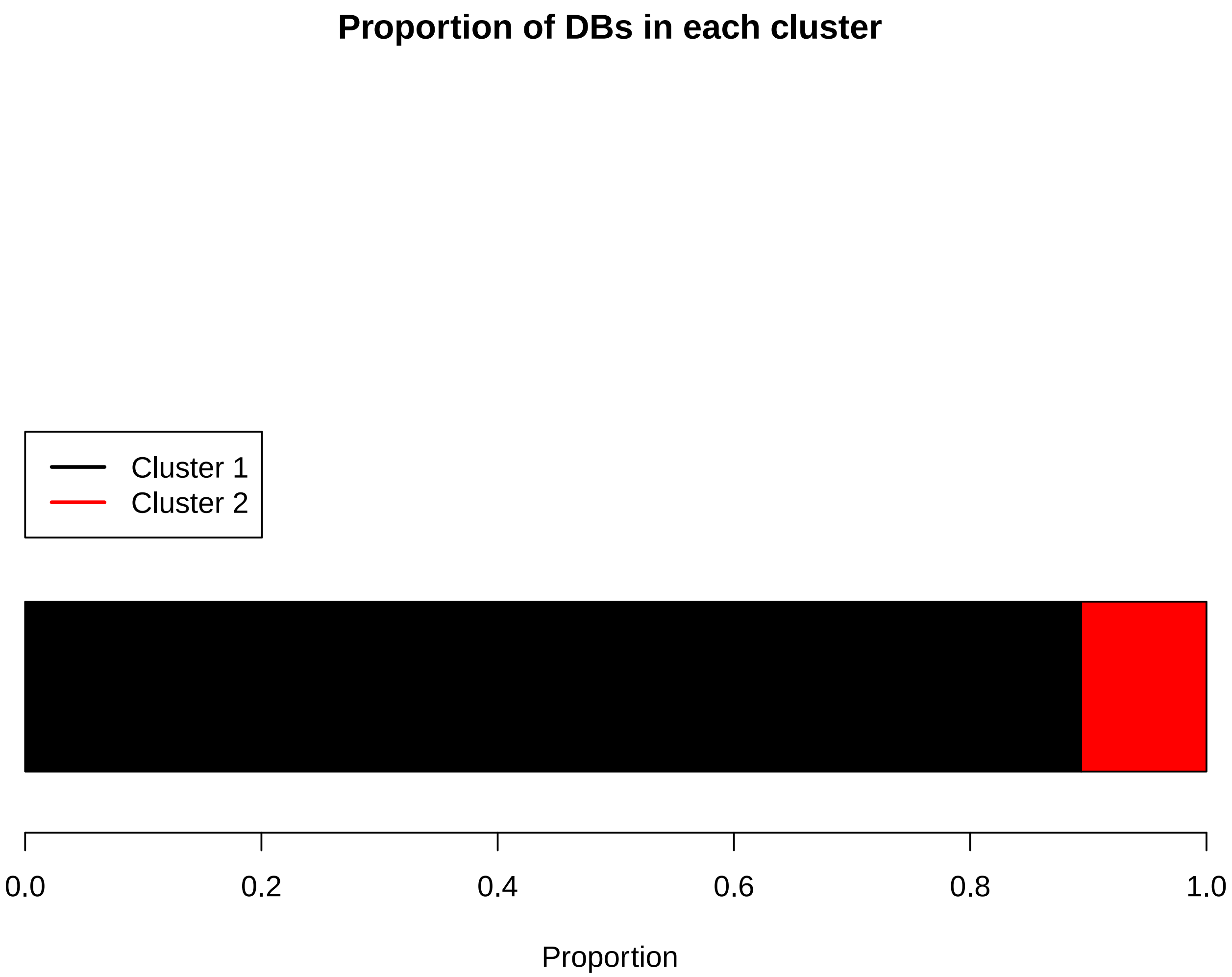
```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2 Cluster 3
##      18120       651      5712
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2 Cluster 3
##       71.8      74.1      70.1
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2 Cluster 3
##     243085  223777.6  233629.8
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2 Cluster 3
##        7871       296      2460
## B      7604       264      2444
## D      1980        71       609
## K       665        20       199
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2 Cluster 3
##      0.226     0.228     0.227
##
##
##
##  Provinces:
##                      Cluster 1 Cluster 2 Cluster 3
## Alberta                    556        17       190
## BritishColumbia            894        31       274
## NewBrunswick               151         5        51
## NorthwestTerritories        11         2         3
## NovaScotia                 579        21       179
## Ontario                   2866       100       892
## Quebec                    1044        29       305
## Saskatchewan                88         3        30
## NA's                     11931       443      3788
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2 Cluster 3
## 0     16352       582      5184
## 1      1364        52       399
## 2       207         7        53
## F       197        10        76
##
##
##
##  PMS_prox_idx_pharma :
##  Cluster 1 Cluster 2 Cluster 3
##    0.04492   0.04294   0.04332
```

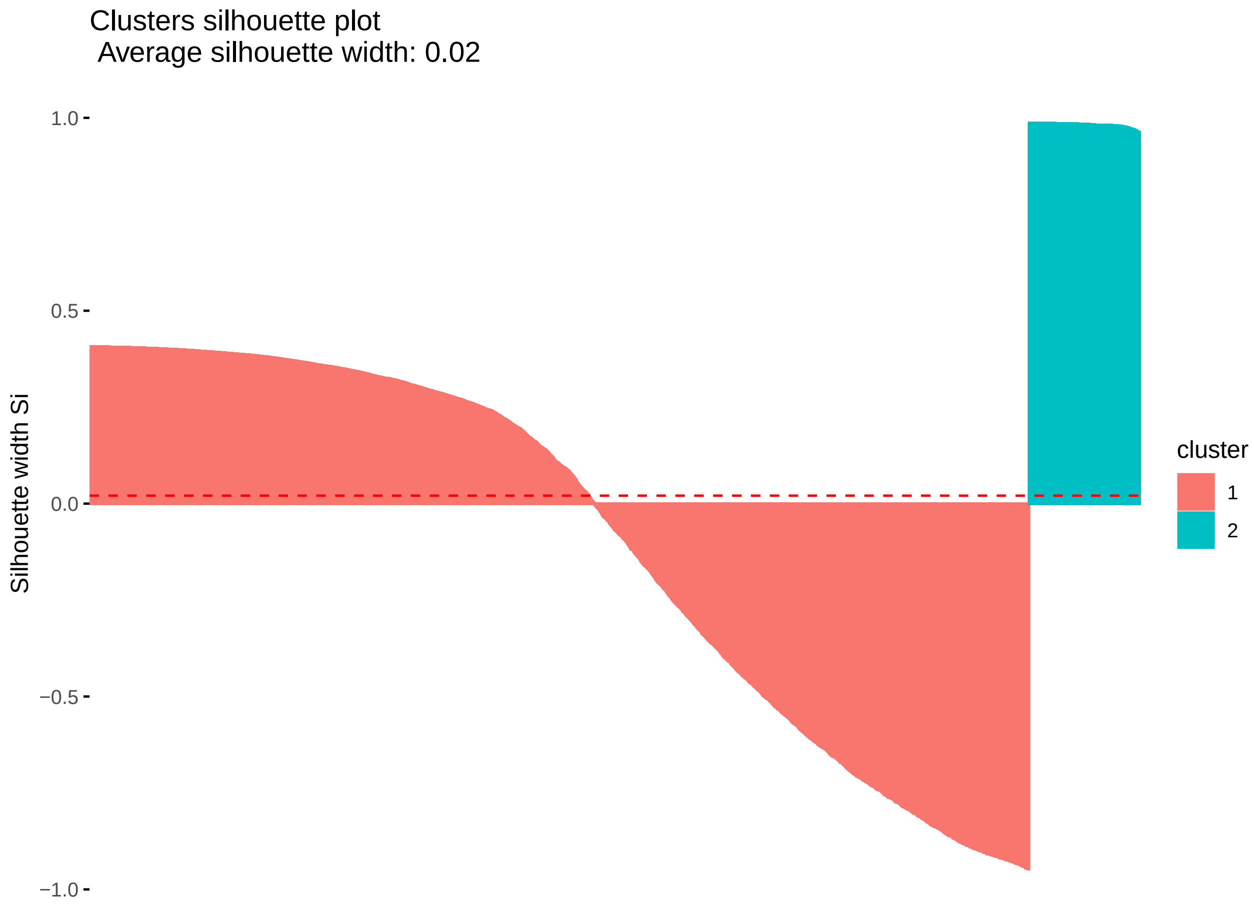
Childcare



```
## [1] "Silhouette coefficient: 0.436142083007771"
## [1] "Xie Beni coefficient: Inf"
## [1] "Davies Bouldin coefficient: 1.77390364670791"
## [1] "Dunn Index coefficient: 0"
## [1] "Calinski-Harabasz coefficient: 3853.86902961714"
```



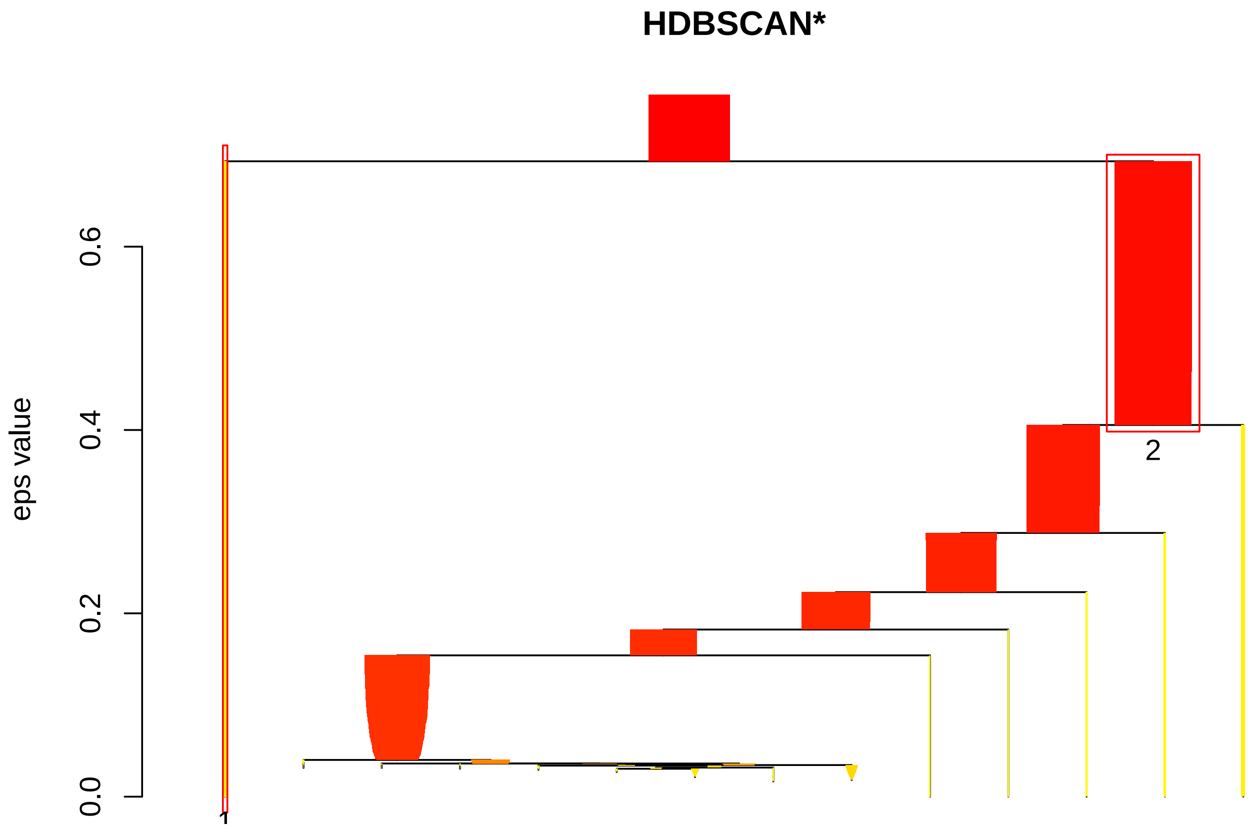
```
## [1] "Segment cutoff values:"
## [1] 0.009
##   cluster  size ave.sil.width
## 1       1 10332         -0.09
## 2       2  1227          0.98
```



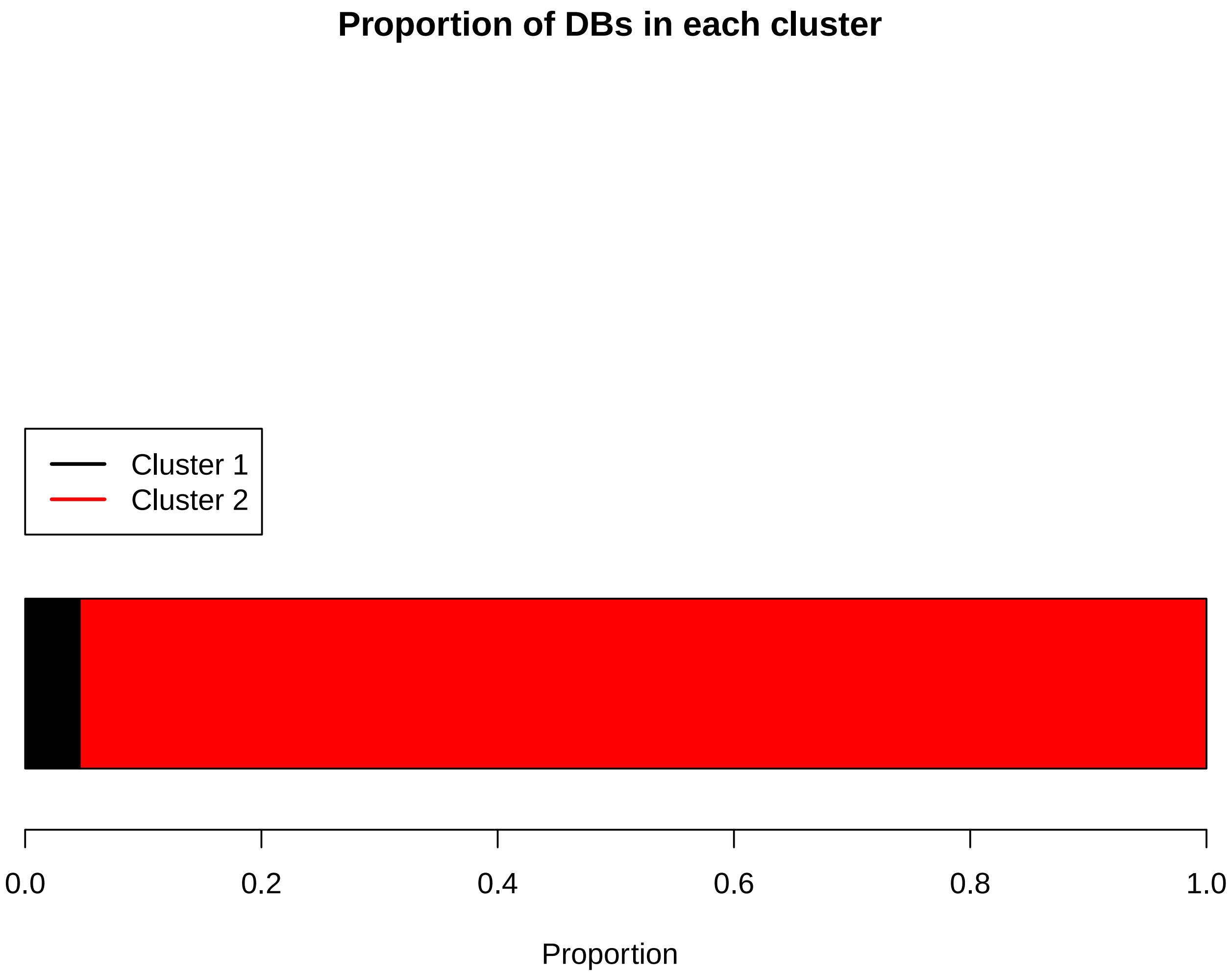
```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##      21897      2586
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       71.6      70.4
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   240617.8  238224.7
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2
##        9485      1142
## B      9234      1078
## D      2393       267
## K       785        99
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2
##      0.226     0.229
##
##
##
##  Provinces:
##                      Cluster 1 Cluster 2
## Alberta                    694        69
## BritishColumbia           1059       140
## NewBrunswick               186        21
## NorthwestTerritories        14         2
## NovaScotia                 689        90
## Ontario                   3464       394
## Quebec                    1215       163
## Saskatchewan               112         9
## NA's                     14464      1698
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0     19785      2333
## 1      1624       191
## 2       238        29
## F       250        33
##
##
##
##  PMS_prox_idx_childcare :
##  Cluster 1 Cluster 2
##    0.07629   0.08111
```

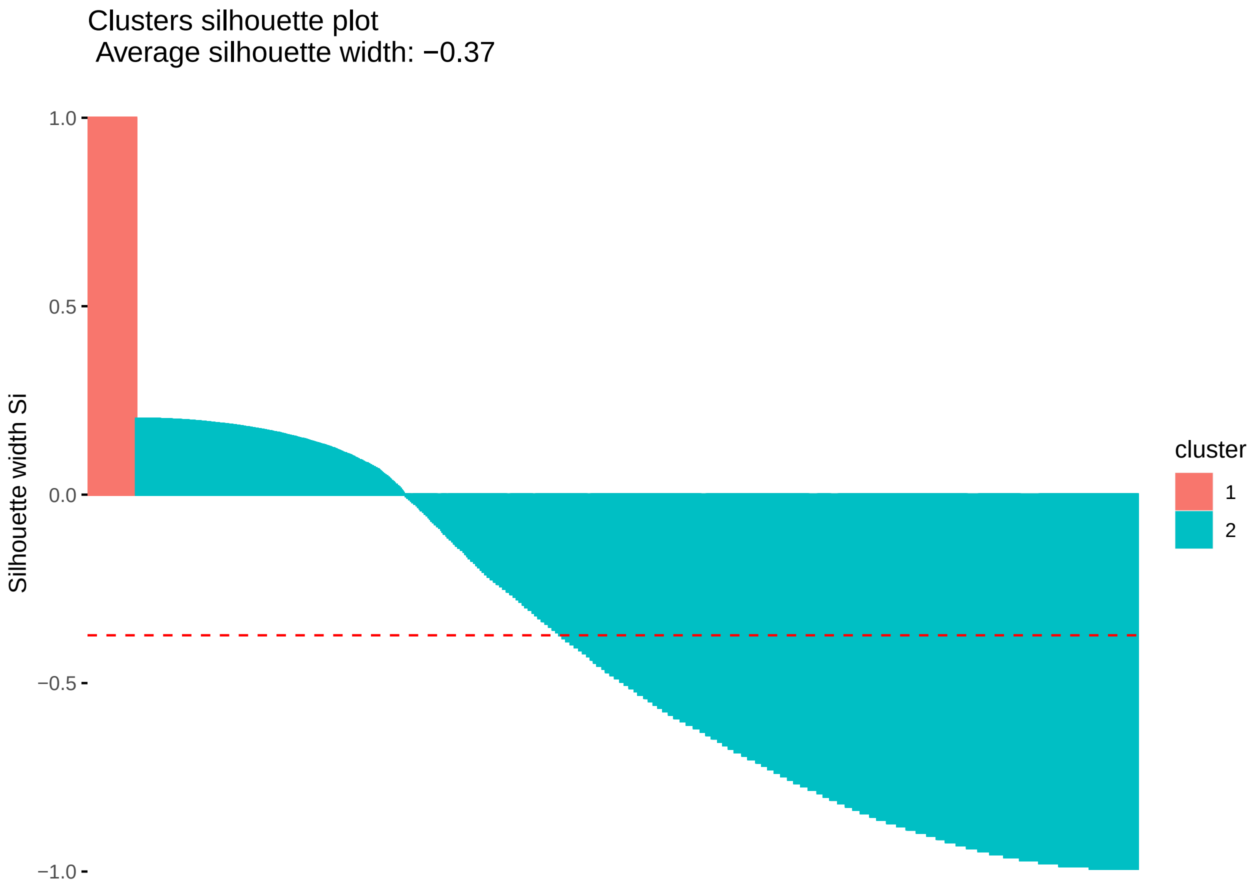
Health



```
## [1] "Silhouette coefficient: 0.732102639194725"
## [1] "Xie Beni coefficient: 5771.3504185073"
## [1] "Davies Bouldin coefficient: 0.348803365749021"
## [1] "Dunn Index coefficient: 0.00290586745518542"
## [1] "Calinski-Harabasz coefficient: 2259.78258147251"
```



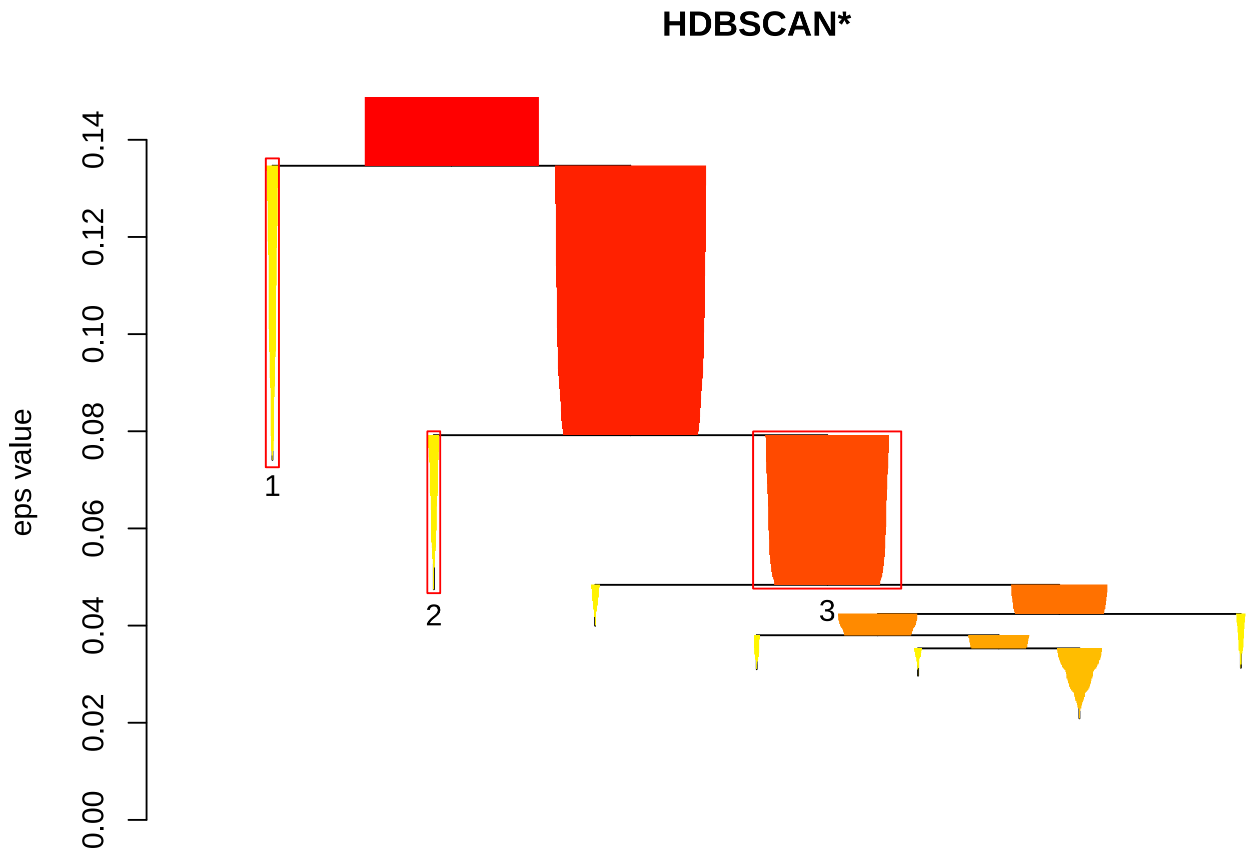
```
## [1] "Segment cutoff values:"
## [1] 0.1052
##   cluster  size ave.sil.width
## 1       1   687          1.00
## 2       2 14191         -0.44
```



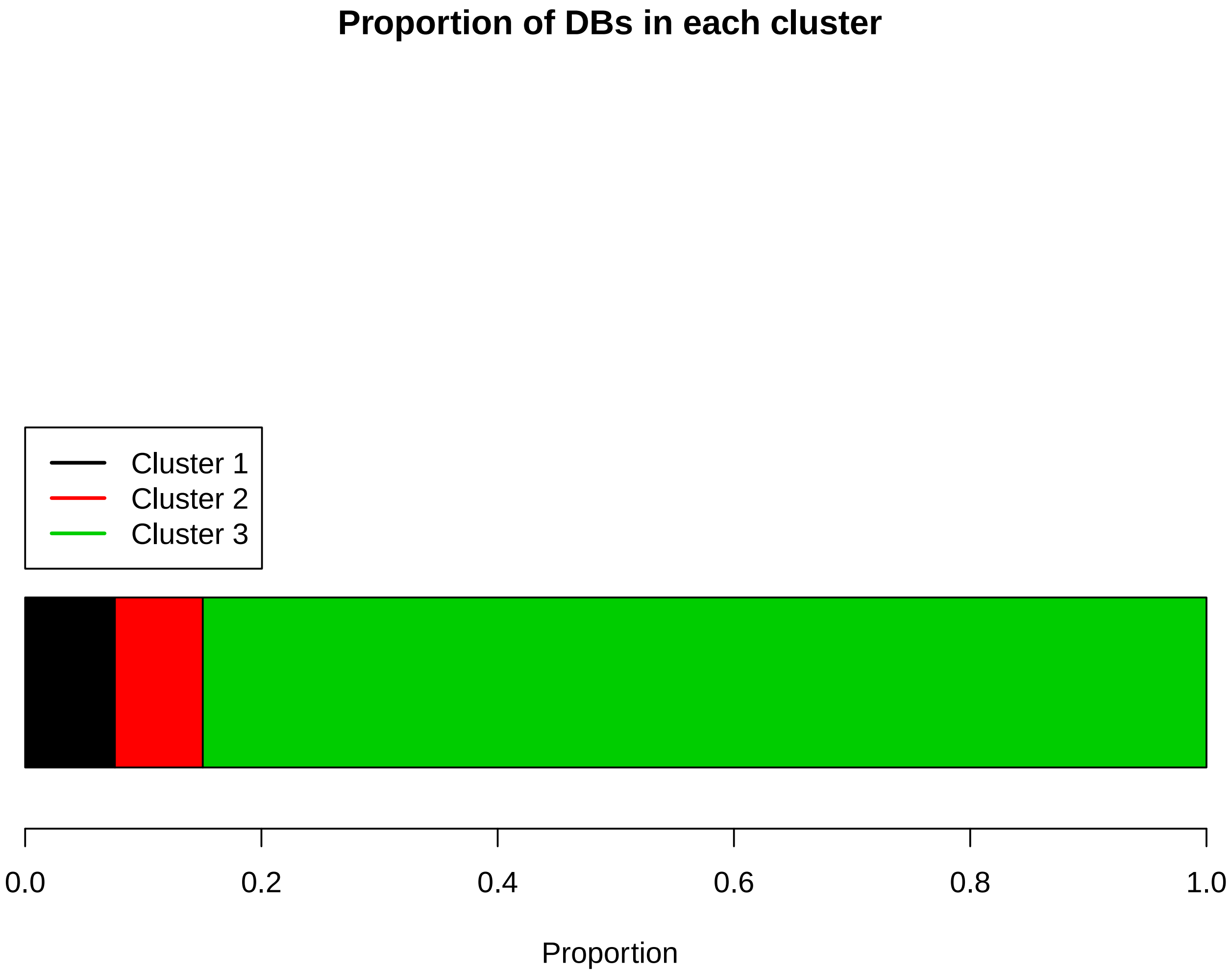
```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##       1138     23345
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       71.2      71.5
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   234132.8  240668.7
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2
##         515     10112
## B       451      9861
## D       128      2532
## K        44       840
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2
##      0.229     0.226
##
##
##
##  Provinces:
##                      Cluster 1 Cluster 2
## Alberta                     31       732
## BritishColumbia             48      1151
## NewBrunswick                14       193
## NorthwestTerritories         0        16
## NovaScotia                  42       737
## Ontario                    177      3681
## Quebec                      62      1316
## Saskatchewan                 6       115
## NA's                       758     15404
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0      1032     21086
## 1        78      1737
## 2        14       253
## F        14       269
##
##
##
##  PMS_prox_idx_health :
##  Cluster 1 Cluster 2
##     0.0138   0.01371
```

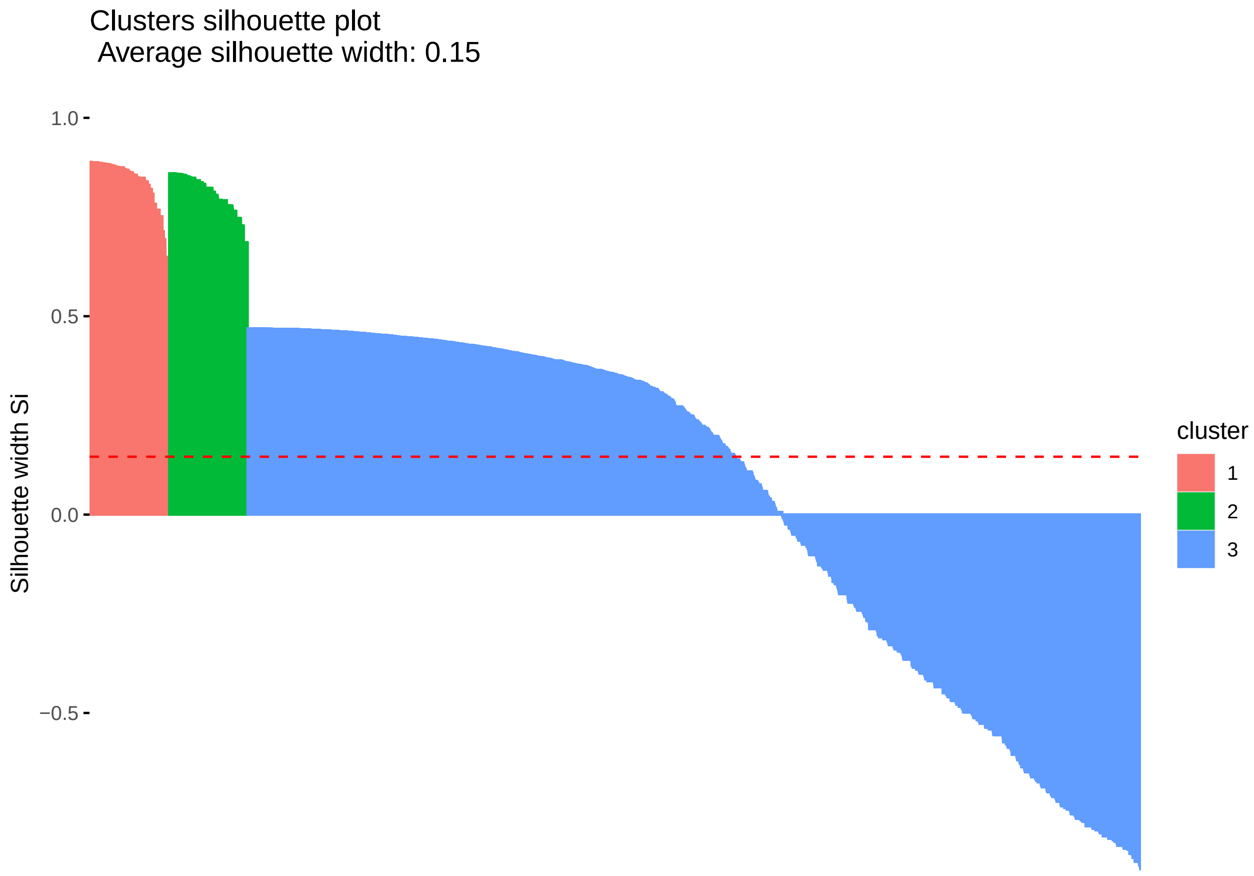
Grocery



```
## [1] "Silhouette coefficient: 0.488393299110873"
## [1] "Xie Beni coefficient: Inf"
## [1] "Davies Bouldin coefficient: 1.15635989806671"
## [1] "Dunn Index coefficient: 0"
## [1] "Calinski-Harabasz coefficient: 1953.01889097893"
```



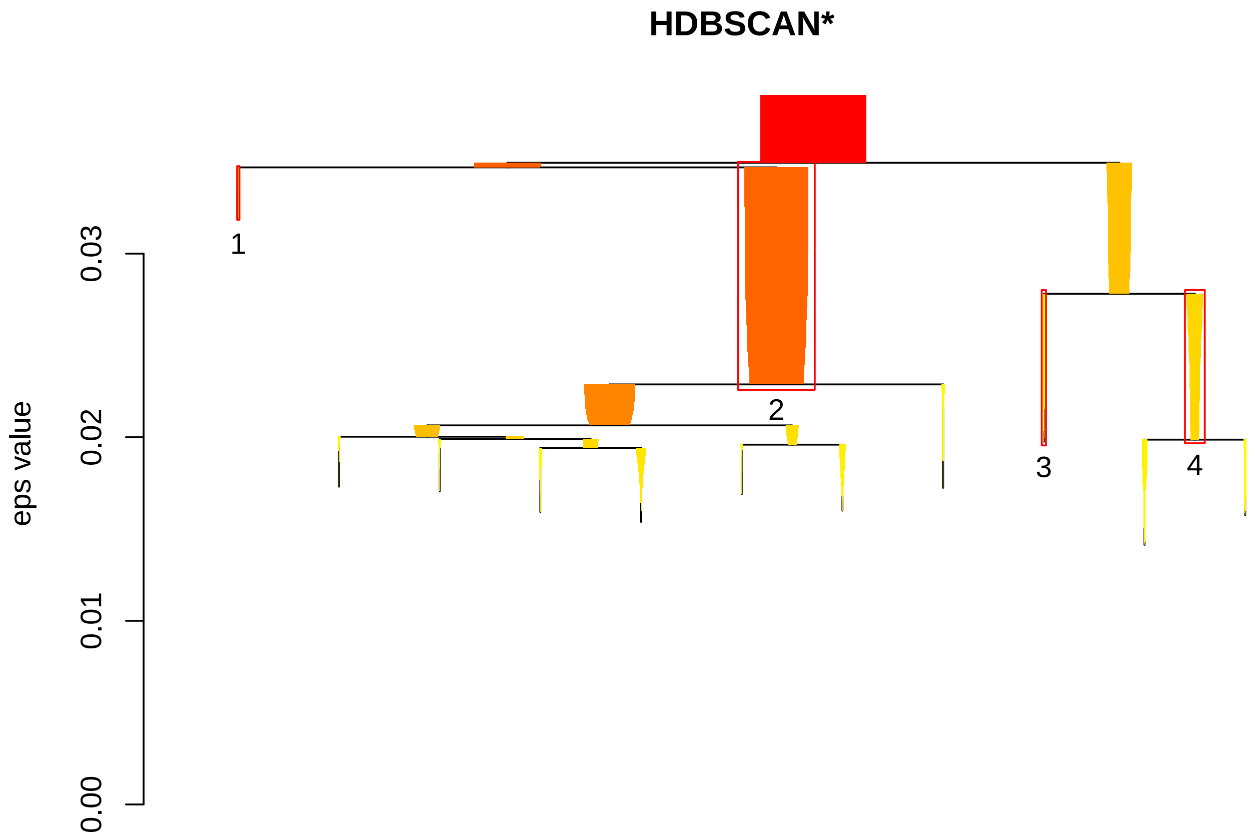
```
## [1] "Segment cutoff values:"
## [1] 0.0763
## [1] 0.0124
##   cluster size ave.sil.width
## 1       1  447          0.84
## 2       2  438          0.81
## 3       3 5001          0.03
```



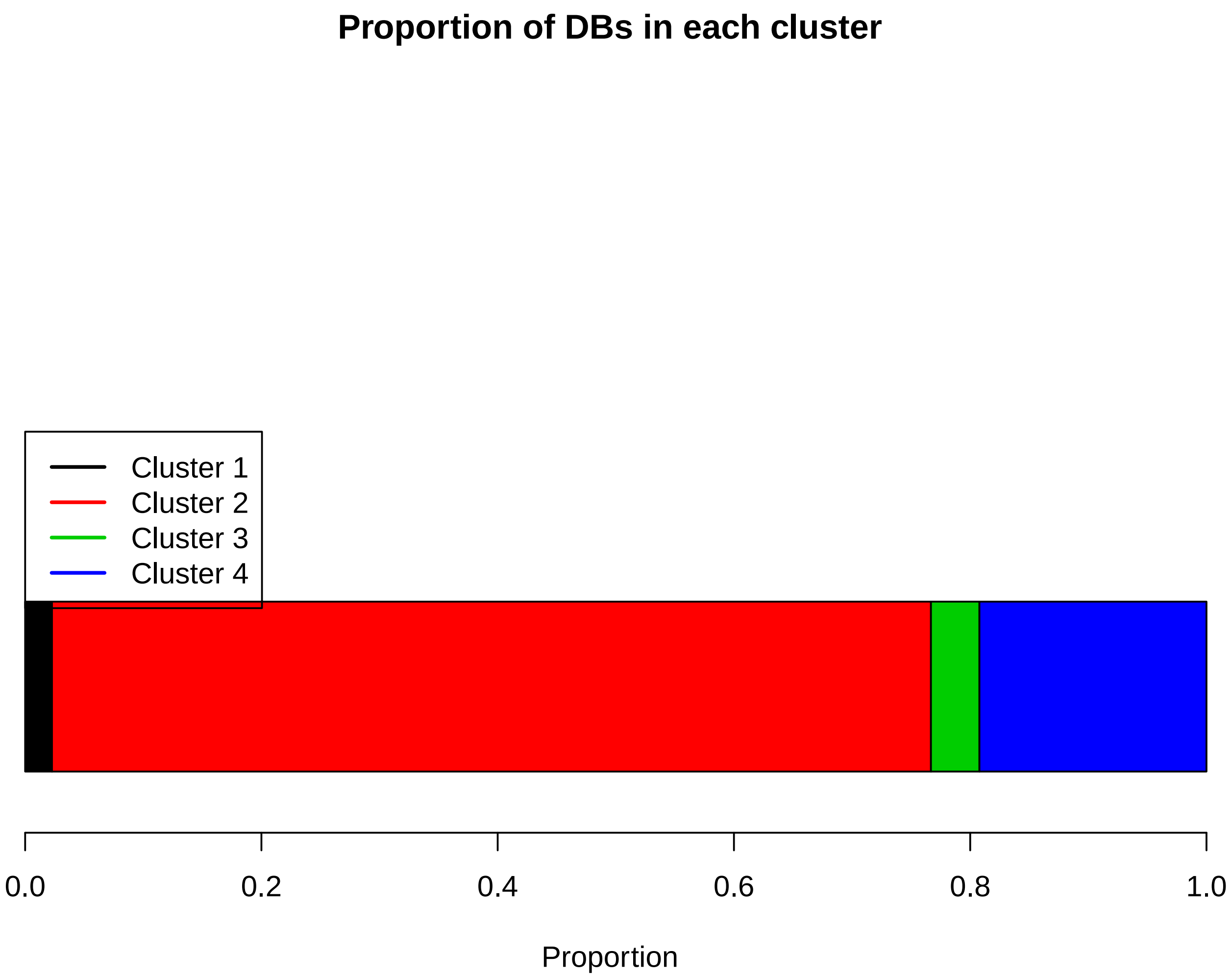
```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2 Cluster 3
##       1870      1812     20801
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2 Cluster 3
##         68      67.7      72.1
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2 Cluster 3
##   231840.4  239585.6  241198.3
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2 Cluster 3
##         854       803      8970
## B       725       745      8842
## D       225       202      2233
## K        66        62       756
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2 Cluster 3
##      0.233     0.233     0.225
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2 Cluster 3
## Alberta                      53        61       649
## BritishColumbia              95        72      1032
## NewBrunswick                 14        17       176
## NorthwestTerritories          1         1        14
## NovaScotia                   70        68       641
## Ontario                     269       265      3324
## Quebec                      116        92      1170
## Saskatchewan                  9        10       102
## NA's                       1243      1226     13693
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2 Cluster 3
## 0      1685      1652     18781
## 1       140       111      1564
## 2        25        26       216
## F        20        23       240
##
##
##
##  PMS_prox_idx_grocery :
##  Cluster 1 Cluster 2 Cluster 3
##    0.07318   0.07302   0.07042
```

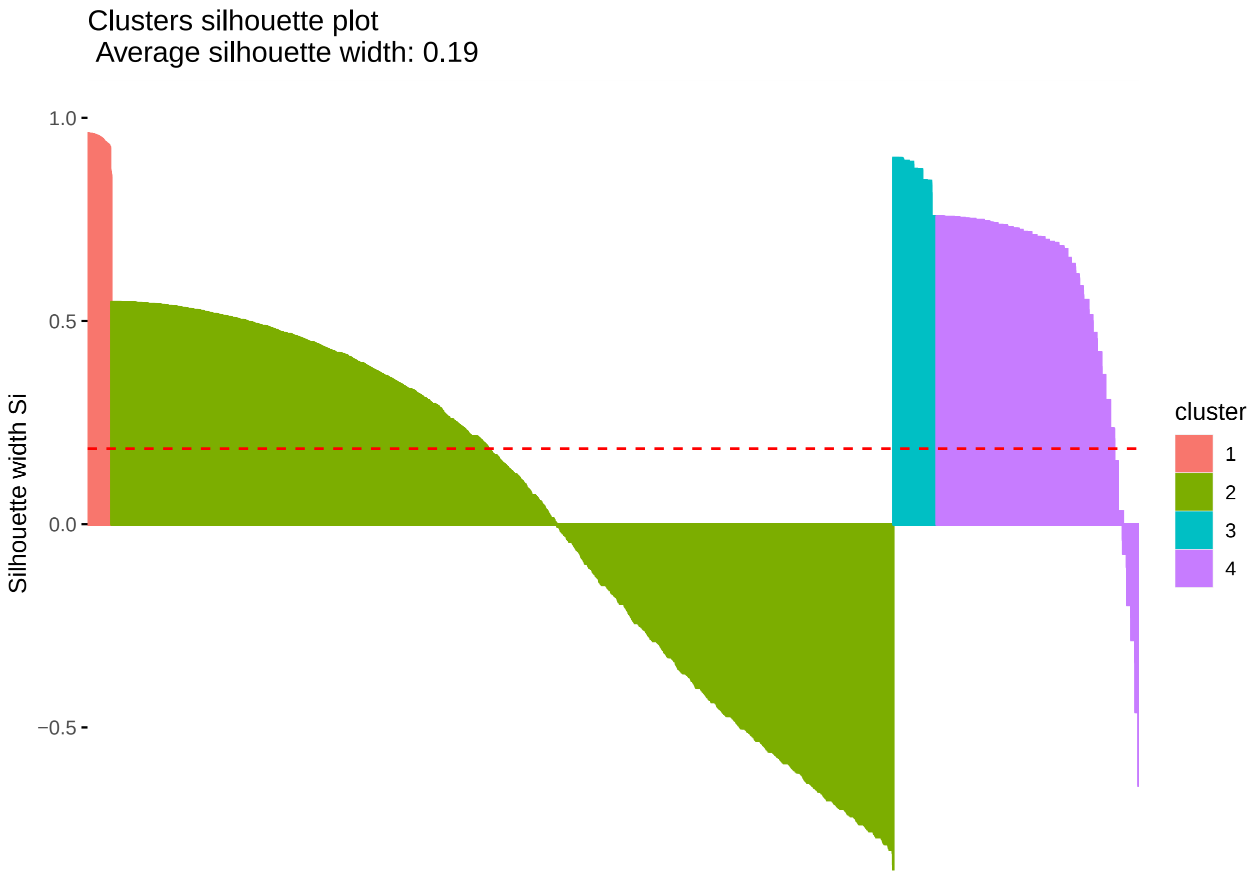
Primary Education



```
## [1] "Silhouette coefficient: 0.332930000531499"
## [1] "Xie Beni coefficient: 1755029.2519329"
## [1] "Davies Bouldin coefficient: 2.69408852023911"
## [1] "Dunn Index coefficient: 8.5279952299452e-05"
## [1] "Calinski-Harabasz coefficient: 2594.06029937711"
```



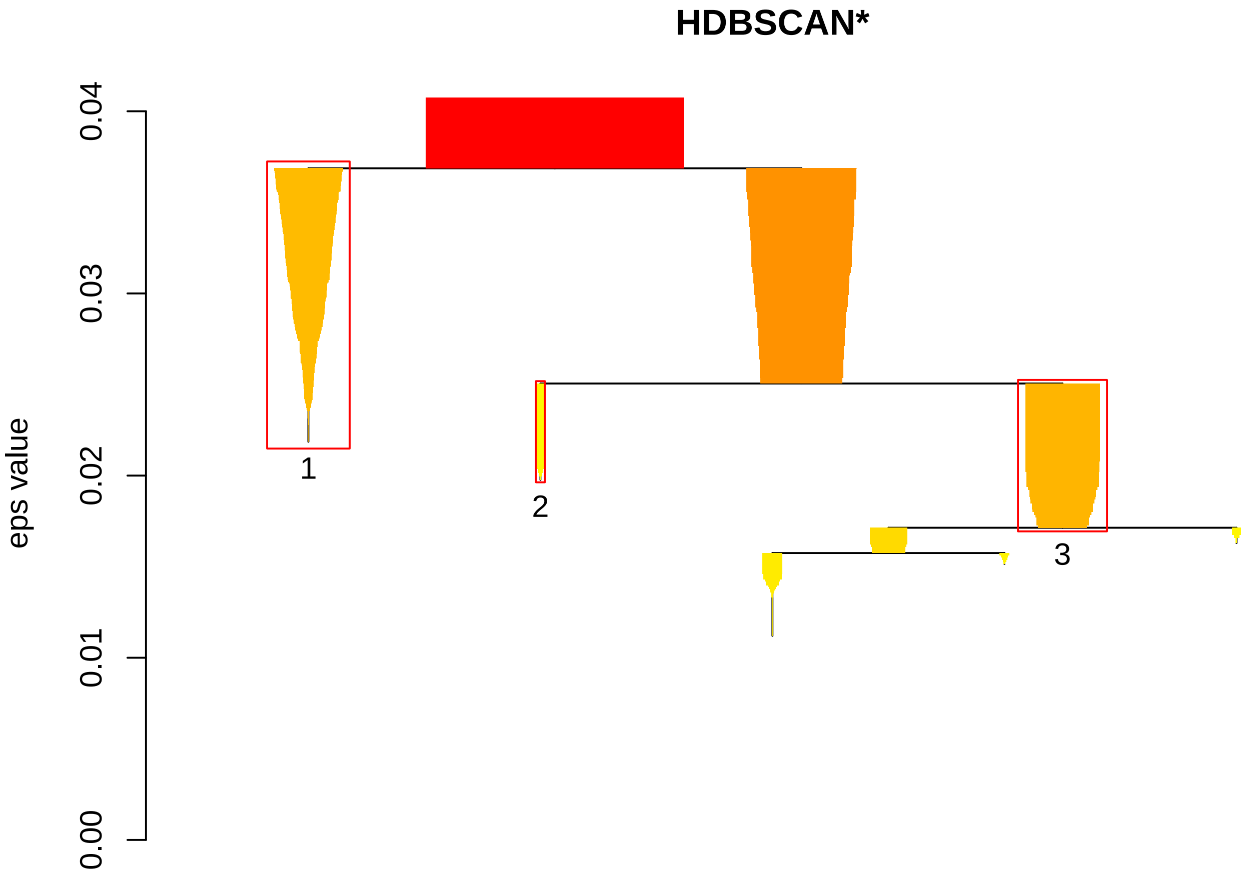
```
## [1] "Segment cutoff values:"
## [1] 0.04495
## [1] 0.22045
## [1] 0.1449
##   cluster size ave.sil.width
## 1       1  208          0.95
## 2       2 6817          0.03
## 3       3  376          0.86
## 4       4 1761          0.57
```



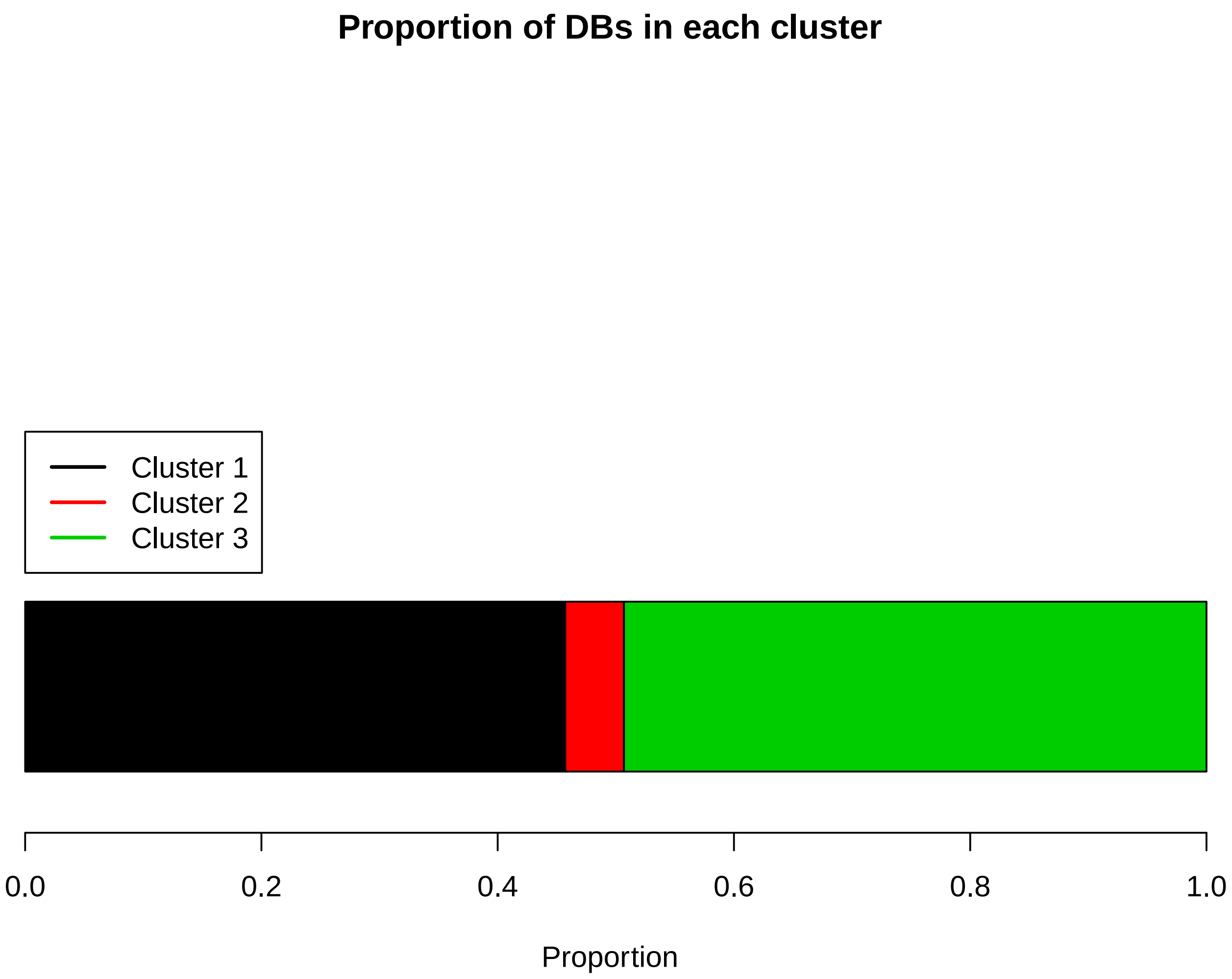
```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##        561     18190      1010      4722
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##       67.7      71.5      66.4      73.1
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##   245090.5  241221.1  255789.7  233207.1
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2 Cluster 3 Cluster 4
##         255      7867       440      2065
## B       233      7690       424      1965
## D        48      1980       114       518
## K        25       653        32       174
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##      0.227     0.226     0.229     0.227
##
##
##
##  Provinces:
##                      Cluster 1 Cluster 2 Cluster 3 Cluster 4
## Alberta                     23       543        35       162
## BritishColumbia             22       892        53       232
## NewBrunswick                 7       159        11        30
## NorthwestTerritories         0        12         2         2
## NovaScotia                  17       575        36       151
## Ontario                     76      2869       167       746
## Quebec                      36      1029        53       260
## Saskatchewan                 1        90         7        23
## NA's                       379     12021       646      3116
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2 Cluster 3 Cluster 4
## 0       520     16399       915      4284
## 1        29      1378        78       330
## 2         3       211        10        43
## F         9       202         7        65
##
##
##
##  PMS_prox_idx_educpri :
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##    0.12297   0.11792   0.11564   0.11525
```

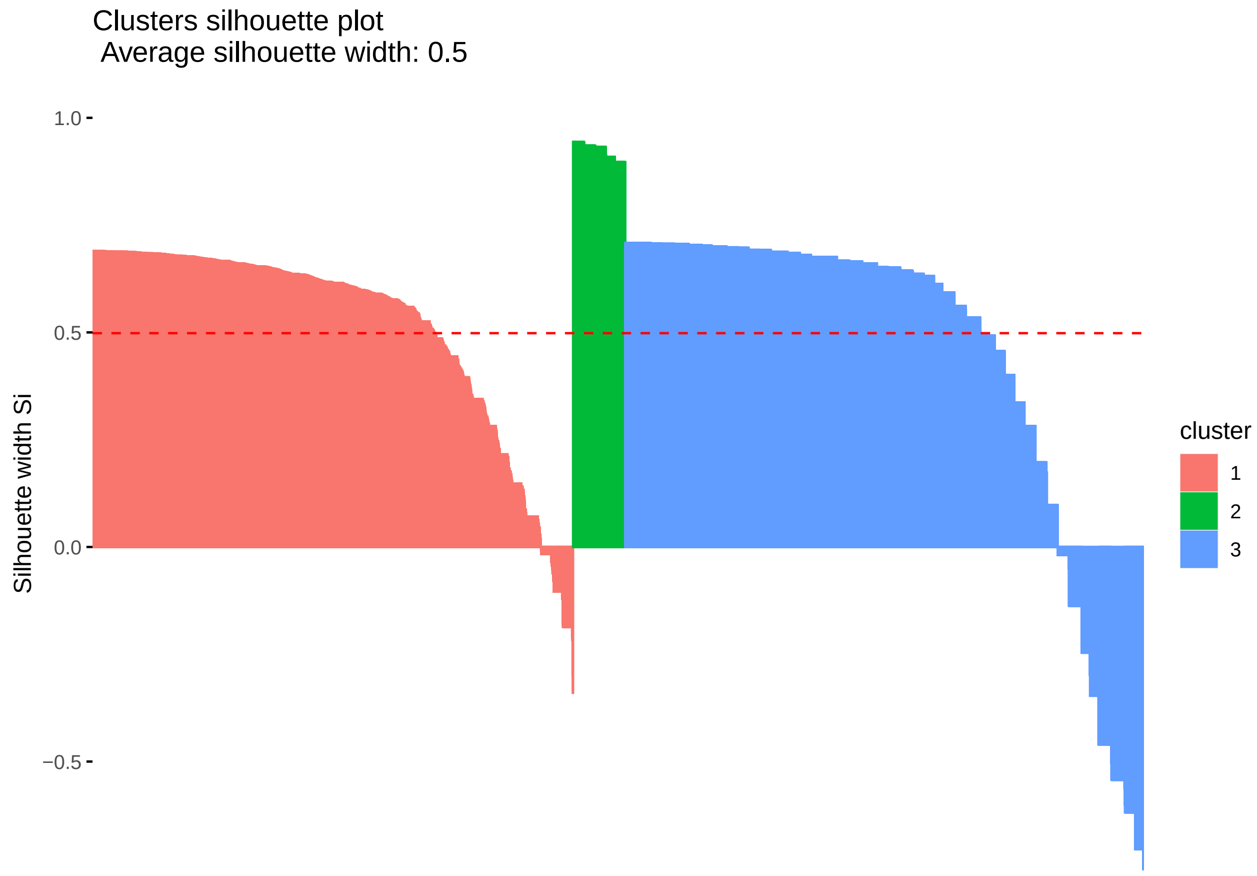
Secondary Education



```
## [1] "Silhouette coefficient: 0.414630345367582"
## [1] "Xie Beni coefficient: 312794.794016636"
## [1] "Davies Bouldin coefficient: 1.3678474078649"
## [1] "Dunn Index coefficient: 0.000181093887703734"
## [1] "Calinski-Harabasz coefficient: 2709.50766404369"
```



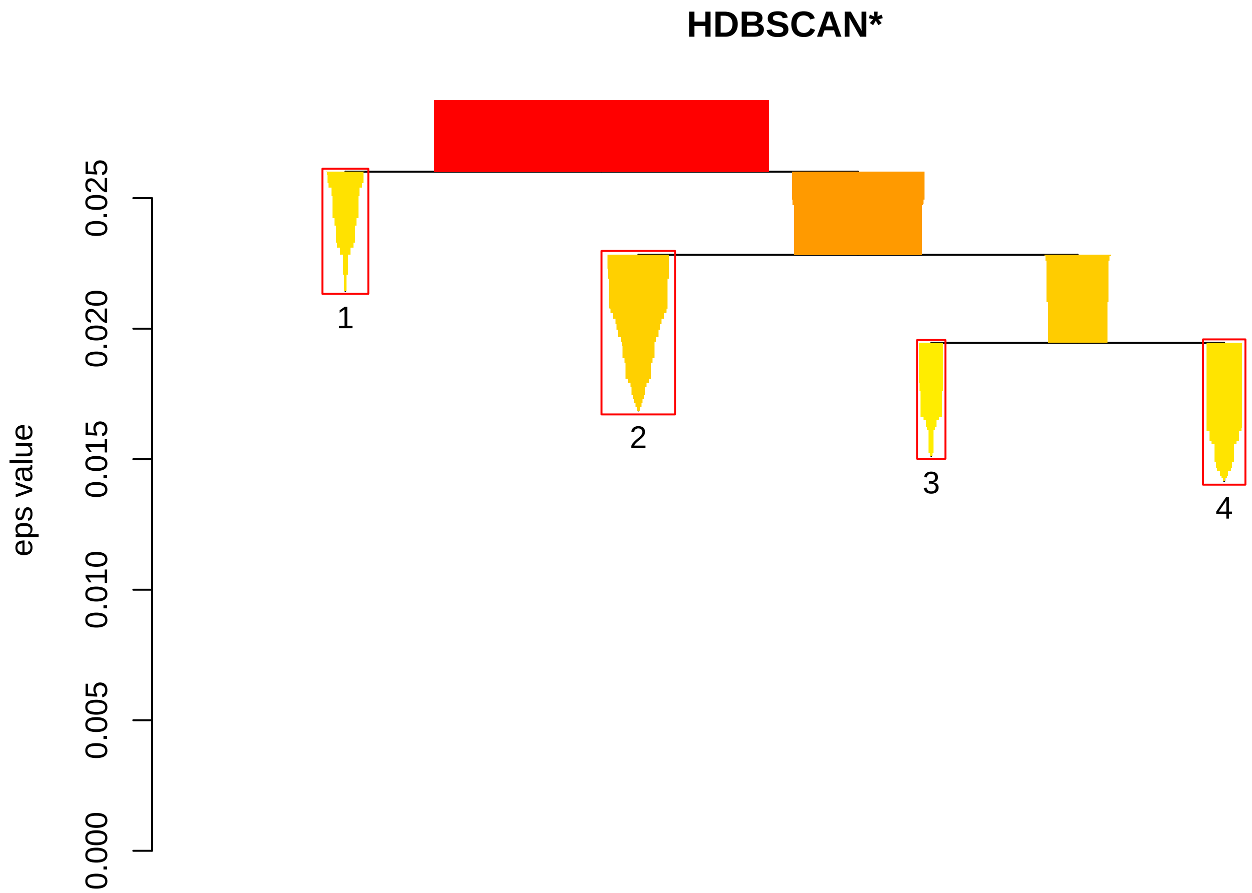
```
## [1] "Segment cutoff values:"
## [1] 0.0576
## [1] 0.0863
##   cluster size ave.sil.width
## 1       1 1886          0.51
## 2       2  205          0.92
## 3       3 2034          0.44
```



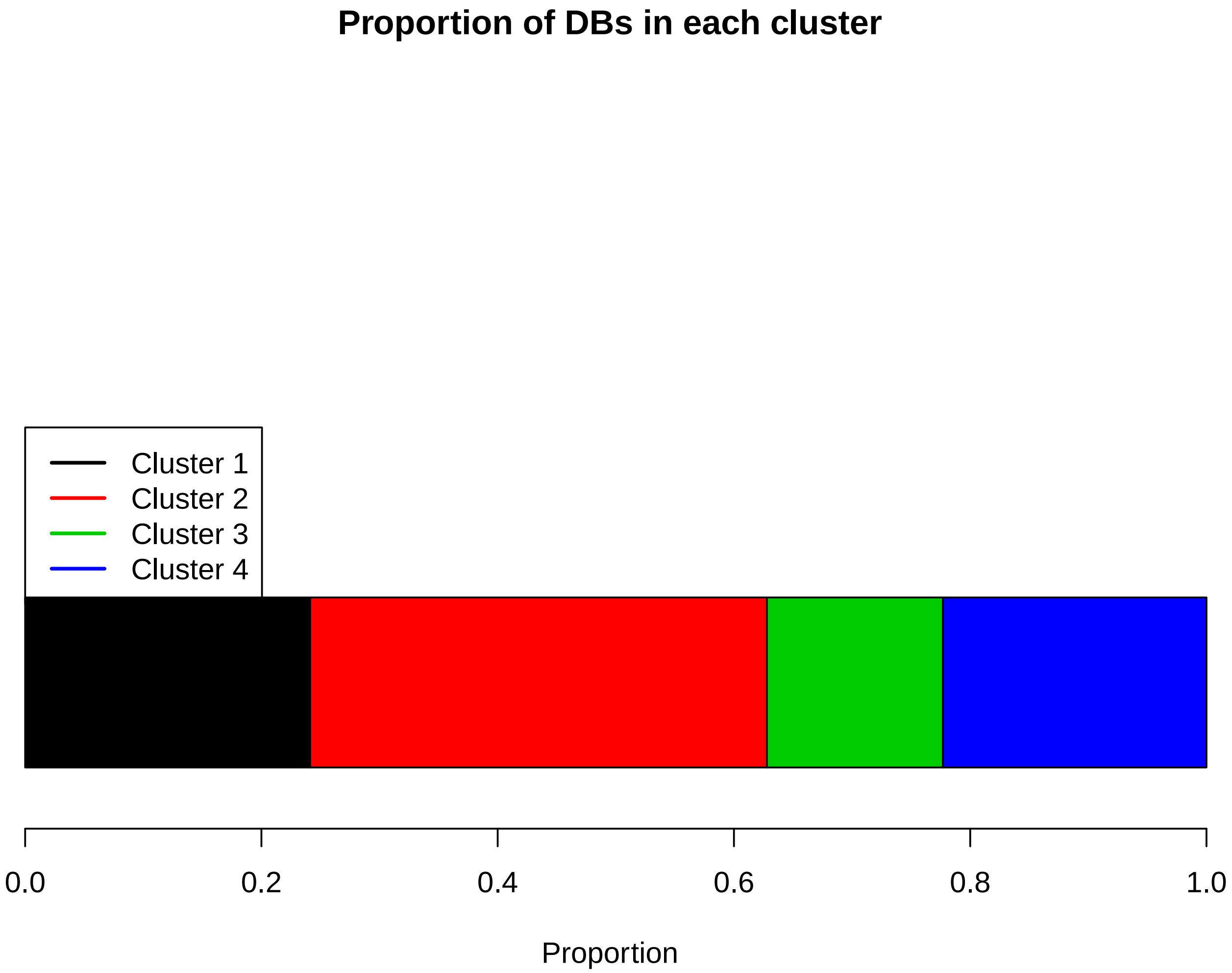
```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2 Cluster 3
##      11200      1207     12076
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2 Cluster 3
##       71.8      76.4      70.6
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2 Cluster 3
##   238827.8  276644.7  238165.6
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2 Cluster 3
##        4799       520      5308
## B      4736       522      5054
## D      1248       126      1286
## K       417        39       428
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2 Cluster 3
##      0.227     0.218     0.227
##
##
##
##  Provinces:
##                      Cluster 1 Cluster 2 Cluster 3
## Alberta                    361        28       374
## BritishColumbia            544        49       606
## NewBrunswick               101         7        99
## NorthwestTerritories         5         0        11
## NovaScotia                 368        41       370
## Ontario                   1743       227      1888
## Quebec                     612        73       693
## Saskatchewan                53         4        64
## NA's                      7413       778      7971
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2 Cluster 3
## 0     10120      1084     10914
## 1       823        94       898
## 2       117        18       132
## F       140        11       132
##
##
##
##  PMS_prox_idx_educsec :
##  Cluster 1 Cluster 2 Cluster 3
##    0.10238   0.10502   0.10468
```

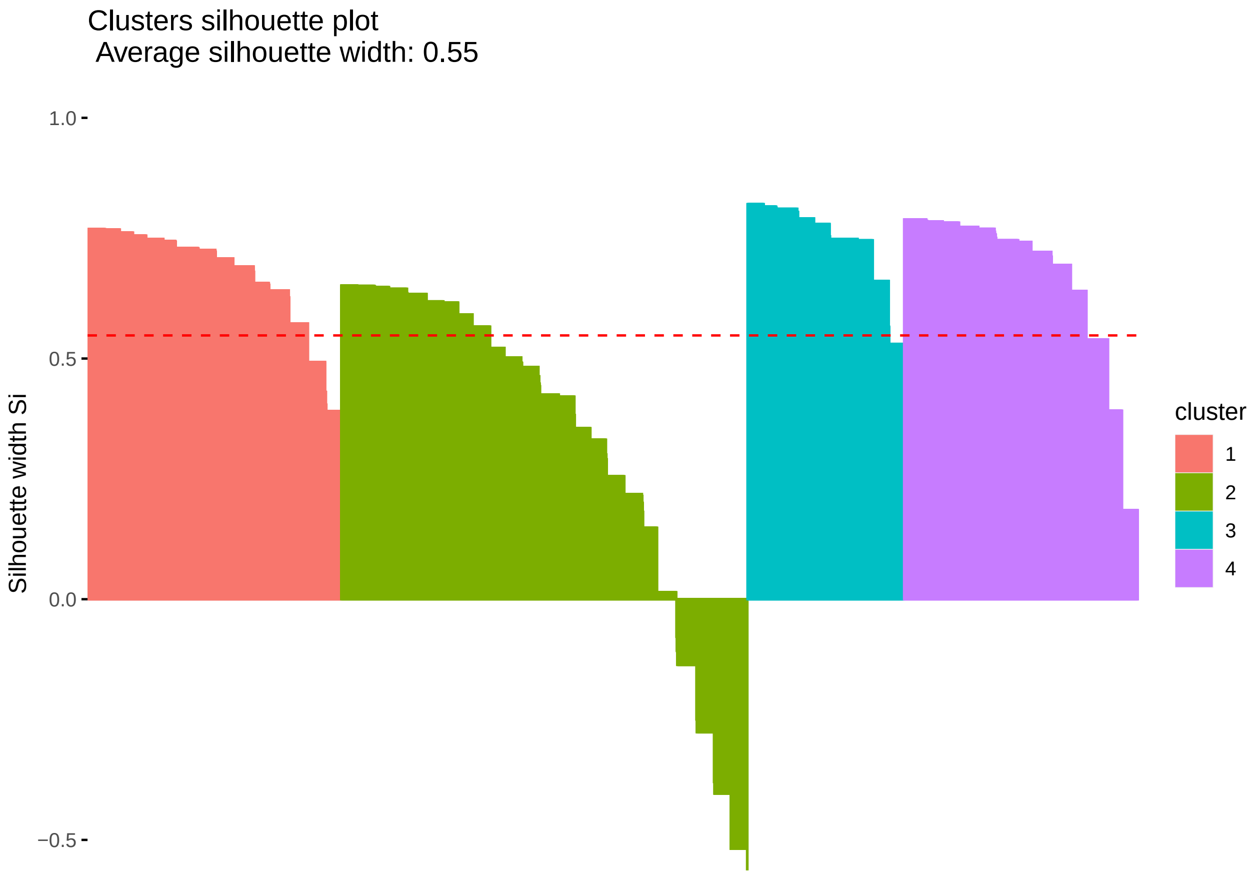
Libraries



```
## [1] "Silhouette coefficient: 0.465481793809224"
## [1] "Xie Beni coefficient: 130260.863229815"
## [1] "Davies Bouldin coefficient: 0.692116879785653"
## [1] "Dunn Index coefficient: 0.000279315860663253"
## [1] "Calinski-Harabasz coefficient: 1394.70607709457"
```



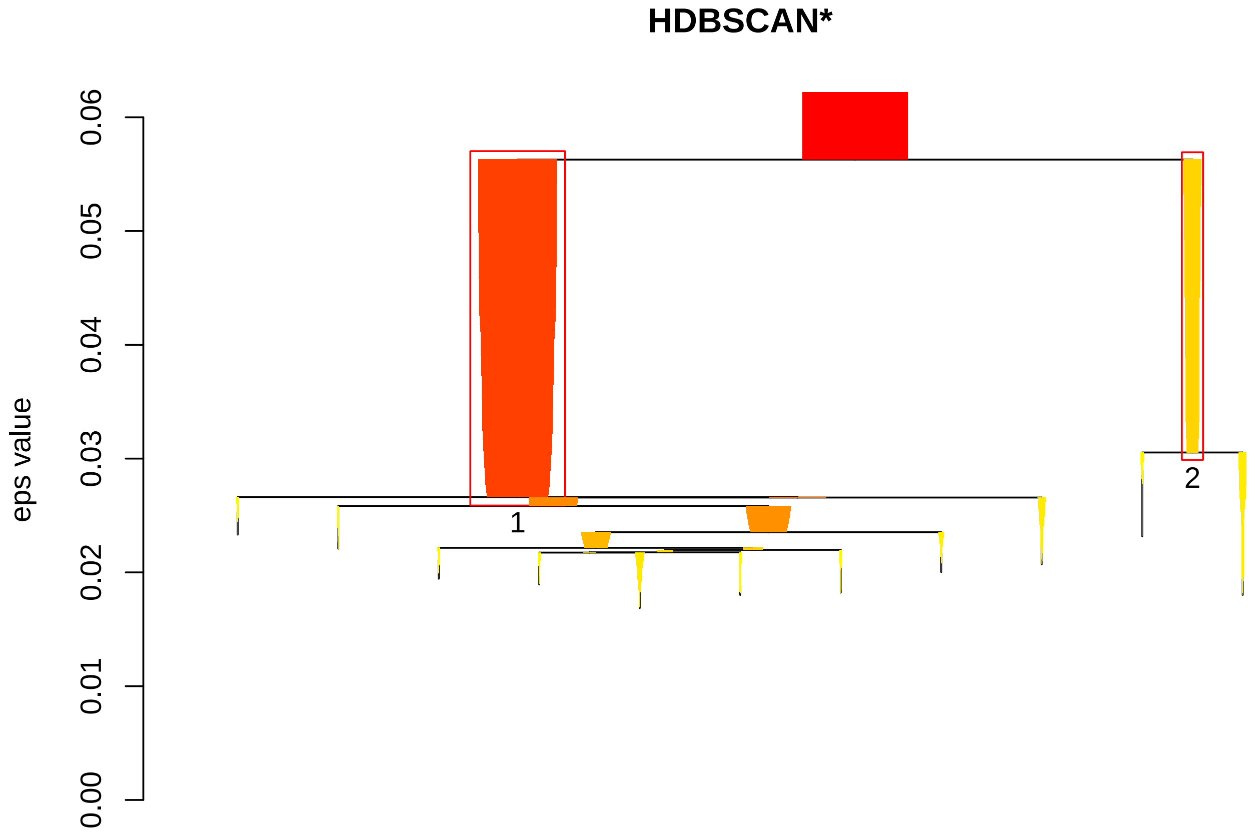
```
## [1] "Segment cutoff values:"
## [1] 0.06575
## [1] 0.0691
## [1] 0.05465
##   cluster size ave.sil.width
## 1       1  635          0.67
## 2       2 1017          0.32
## 3       3  392          0.75
## 4       4  587          0.67
```



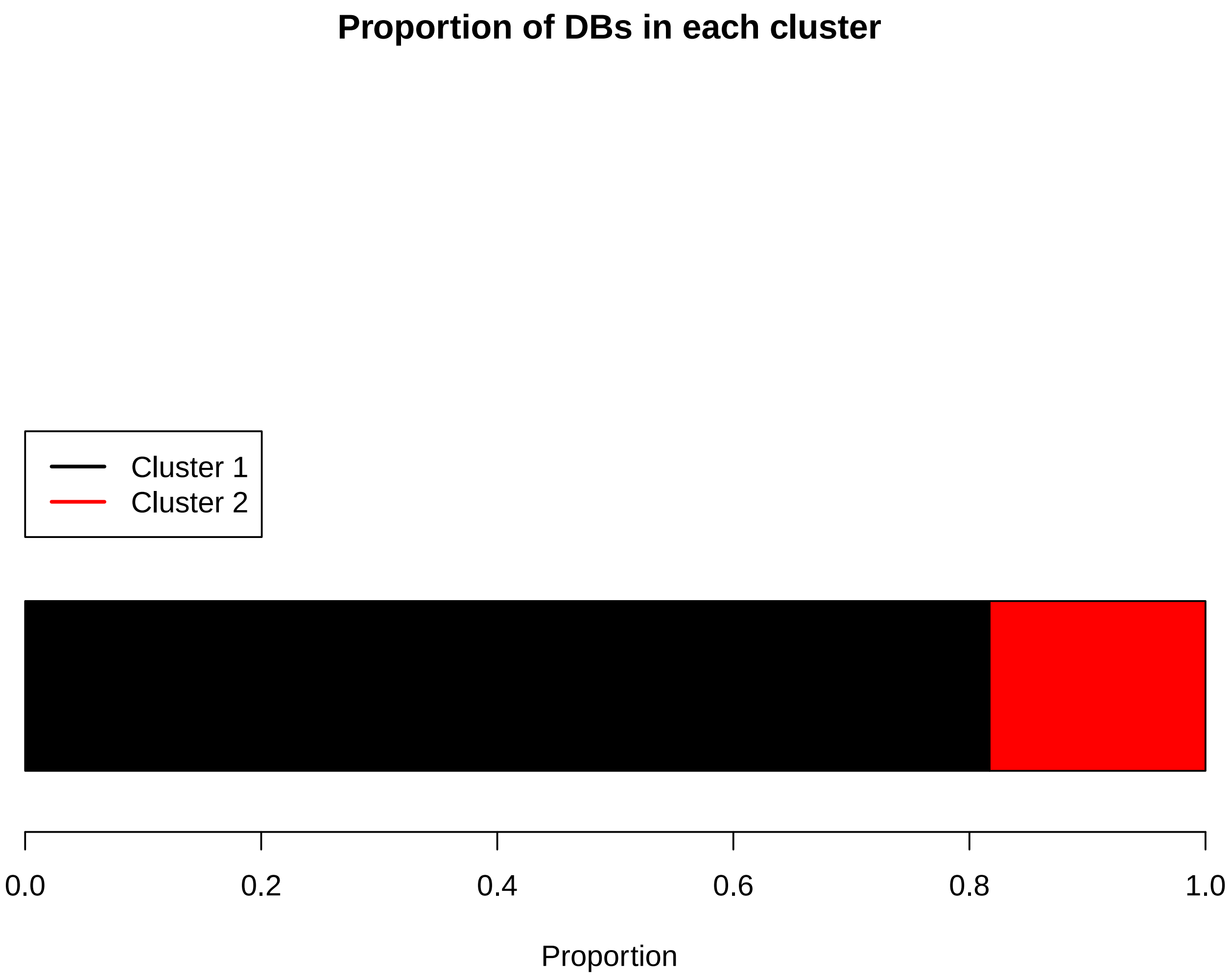
```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##       5907      9464      3648      5464
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##       71.6      73.9      70.9      67.4
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##   239171.8  238945.8  255939.7  233711.8
##
##
##
##  CMA Type:
##    Cluster 1 Cluster 2 Cluster 3 Cluster 4
##         2539      4100      1573      2415
## B       2543      3900      1569      2300
## D        606      1098       396       560
## K        219       366       110       189
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##      0.226     0.227     0.224     0.226
##
##
##
##  Provinces:
##                      Cluster 1 Cluster 2 Cluster 3 Cluster 4
## Alberta                    184       291       110       178
## BritishColumbia            274       499       178       248
## NewBrunswick                42        83        33        49
## NorthwestTerritories         3         9         0         4
## NovaScotia                 177       339       106       157
## Ontario                    934      1443       594       887
## Quebec                     343       533       203       299
## Saskatchewan                21        46        17        37
## NA's                      3929      6221      2407      3605
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2 Cluster 3 Cluster 4
## 0      5307      8582      3281      4948
## 1       471       688       271       385
## 2        63        96        46        62
## F        66        98        50        69
##
##
##
##  PMS_prox_idx_lib :
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##    0.11305   0.11576    0.1125   0.11283
```

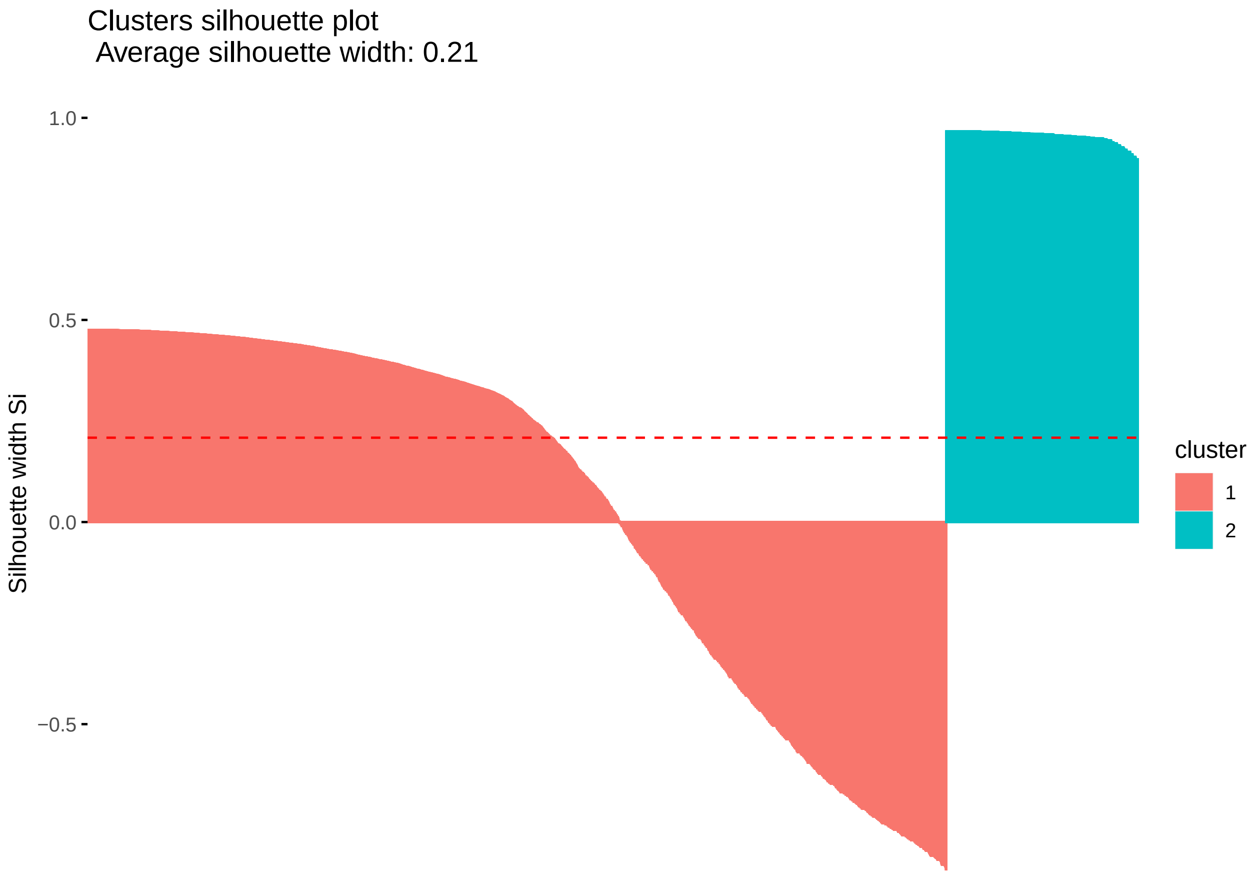
Parks



```
## [1] "Silhouette coefficient: 0.356977727090723"
## [1] "Xie Beni coefficient: Inf"
## [1] "Davies Bouldin coefficient: 4.05632596932578"
## [1] "Dunn Index coefficient: 0"
## [1] "Calinski-Harabasz coefficient: 4007.92463660904"
```



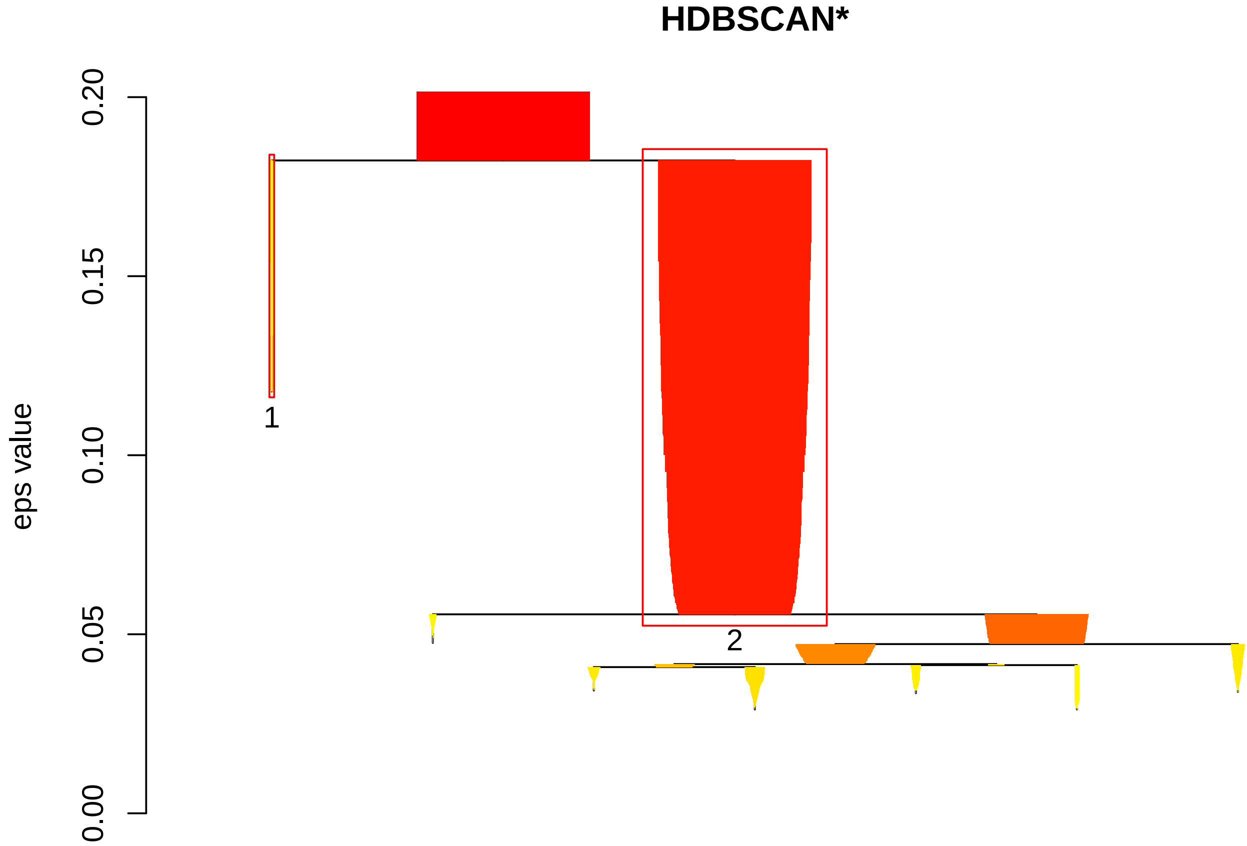
```
## [1] "Segment cutoff values:"
## [1] 0.01995
##   cluster size ave.sil.width
## 1       1 8769          0.04
## 2       2 1963          0.95
```


Clusters silhouette plot
 Average silhouette width: 0.21

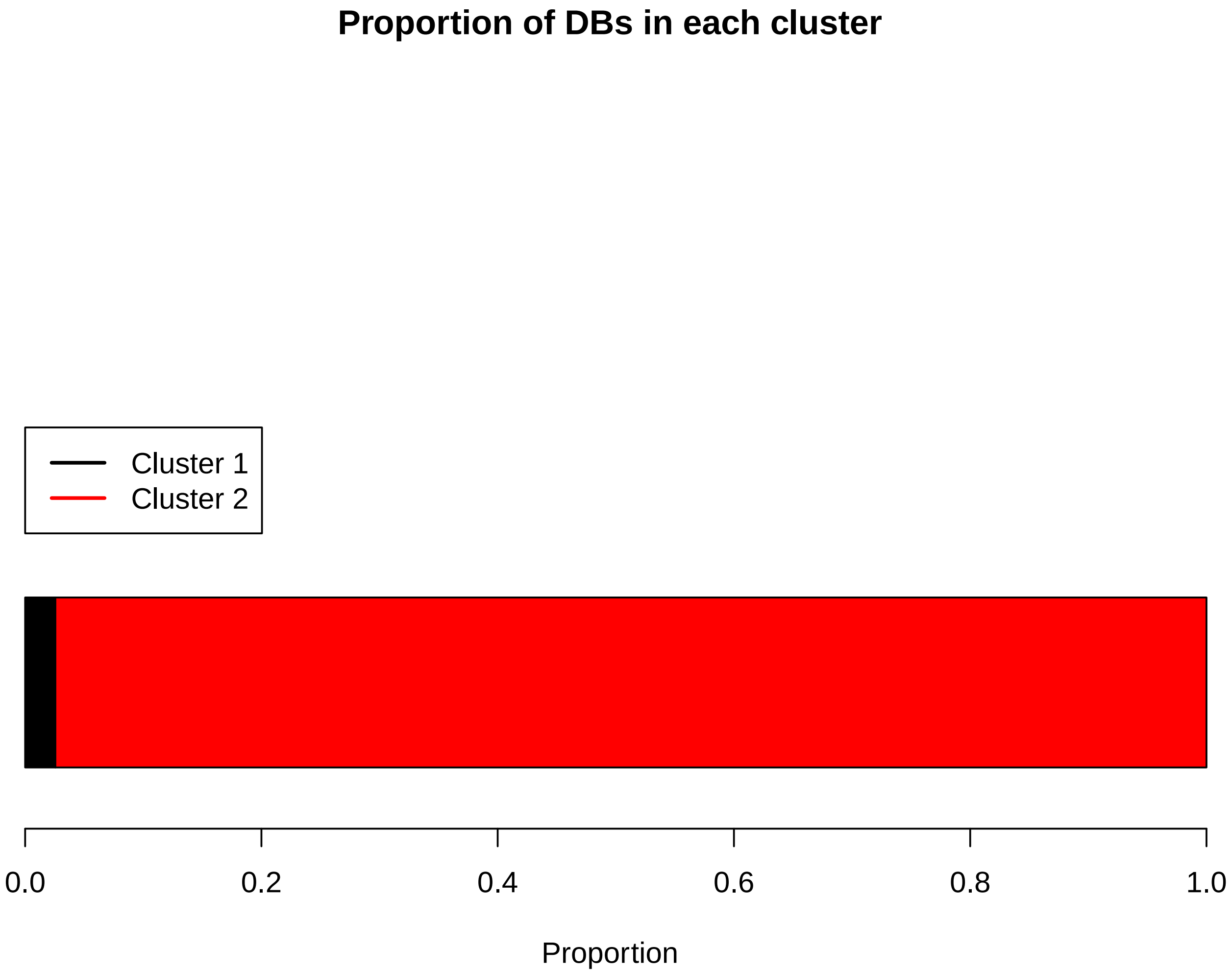
```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##      20024      4459
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       70.7      74.9
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   236787.4  256429.3
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2
##        8737      1890
## B      8415      1897
## D      2151       509
## K       721       163
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2
##      0.227     0.223
##
##
##
##  Provinces:
##                      Cluster 1 Cluster 2
## Alberta                    614       149
## BritishColumbia            973       226
## NewBrunswick               173        34
## NorthwestTerritories        13         3
## NovaScotia                 634       145
## Ontario                   3124       734
## Quebec                    1125       253
## Saskatchewan                98        23
## NA's                     13270      2892
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0     18124      3994
## 1      1456       359
## 2       209        58
## F       235        48
##
##
##
##  PMS_prox_idx_parks :
##  Cluster 1 Cluster 2
##    0.06838    0.0697
```

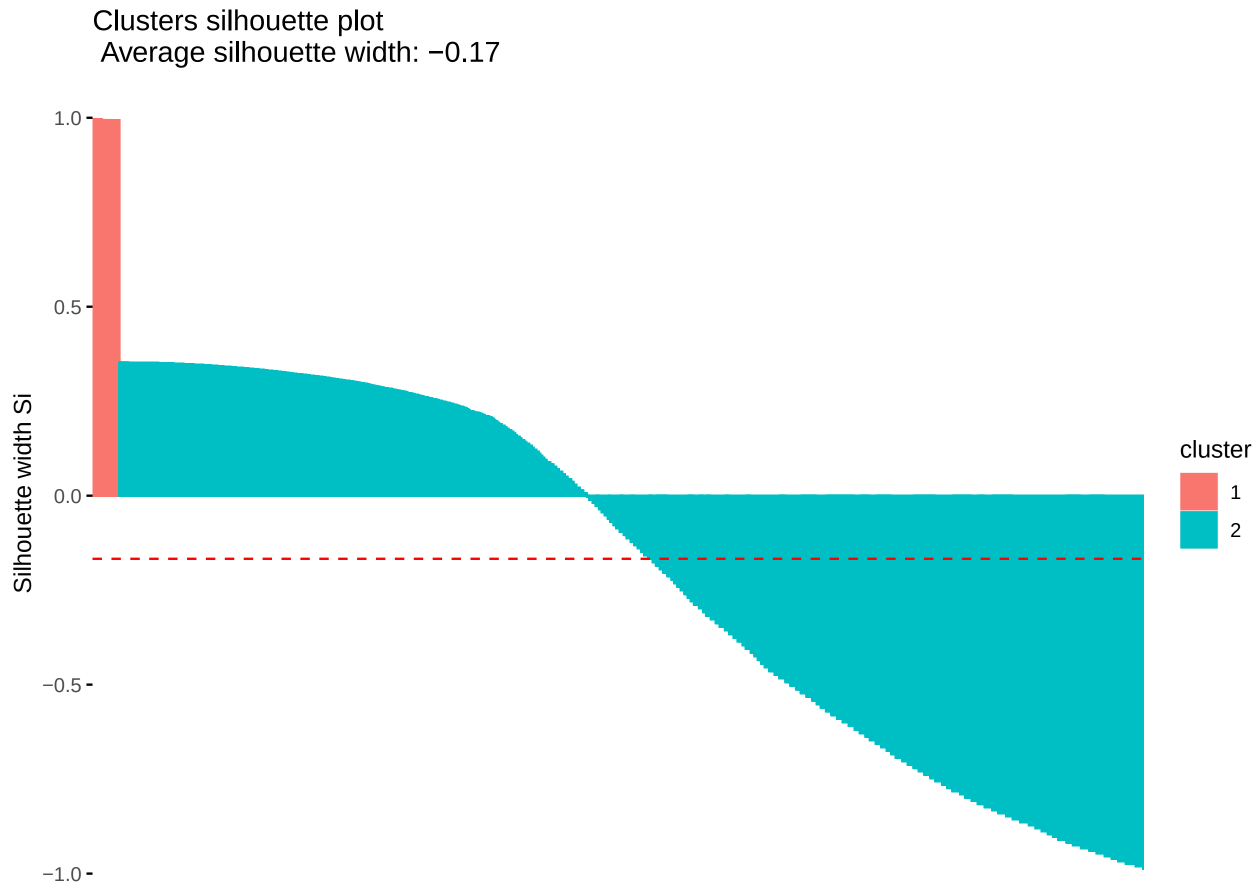
Transit



```
## [1] "Silhouette coefficient: 0.272397309613473"
## [1] "Xie Beni coefficient: Inf"
## [1] "Davies Bouldin coefficient: 2.4566378939336"
## [1] "Dunn Index coefficient: 0"
## [1] "Calinski-Harabasz coefficient: 957.722323613857"
```



```
## [1] "Segment cutoff values:"
## [1] 0.03875
##   cluster size ave.sil.width
## 1       1  210          0.99
## 2       2 7999         -0.20
```



```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##        623     23860
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       58.7      71.8
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   252060.9  240059.4
##
##
##
##  CMA Type:
##    Cluster 1 Cluster 2
##          265     10362
## B        267     10045
## D         63      2597
## K         28       856
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2
##      0.221     0.226
##
##
##
##  Provinces:
##                      Cluster 1 Cluster 2
## Alberta                     18       745
## BritishColumbia             30      1169
## NewBrunswick                 5       202
## NorthwestTerritories         0        16
## NovaScotia                  16       763
## Ontario                    103      3755
## Quebec                      30      1348
## Saskatchewan                 5       116
## NA's                       416     15746
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0       569     21549
## 1        44      1771
## 2         2       265
## F         8       275
##
##
##
##  PMS_prox_idx_transit :
##  Cluster 1 Cluster 2
##    0.01771   0.01809
```

Conclusion

text

```
## list(PMS_prox_idx_emp = 0.22985, PMS_prox_idx_pharma = c(0.01175,
## 0.0525), PMS_prox_idx_childcare = 0.009, PMS_prox_idx_health = 0.1052,
##     PMS_prox_idx_grocery = c(0.0763, 0.0124), PMS_prox_idx_educpri = c(0.04495,
##     0.22045, 0.1449), PMS_prox_idx_educsec = c(0.0576, 0.0863
##     ), PMS_prox_idx_lib = c(0.06575, 0.0691, 0.05465), PMS_prox_idx_parks = 0.01995,
##     PMS_prox_idx_transit = 0.03875)
```