k-means with Imputation

`ClustImpute` package

PMS

17 May, 2023

Assumptions of the Alogrithm

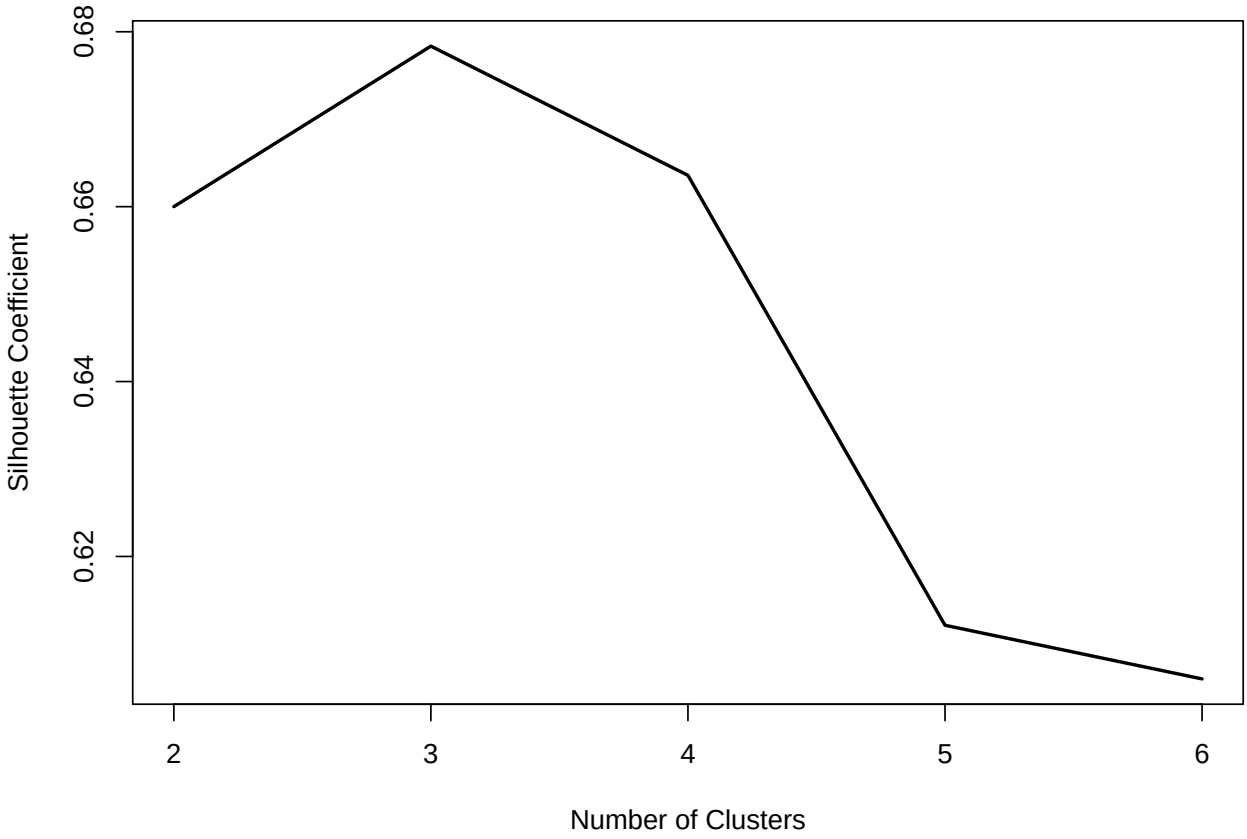
This algorithm "draws the missing values iteratively based on the current cluster assignment so that correlations are considered on this level". Also, "penalizing weights are imposed on imputed values and successively decreased (to zero) as the missing data imputation gets better". The idea is that the missing value is imputed by those other observations that are more similar to it (ie. in the same cluster).

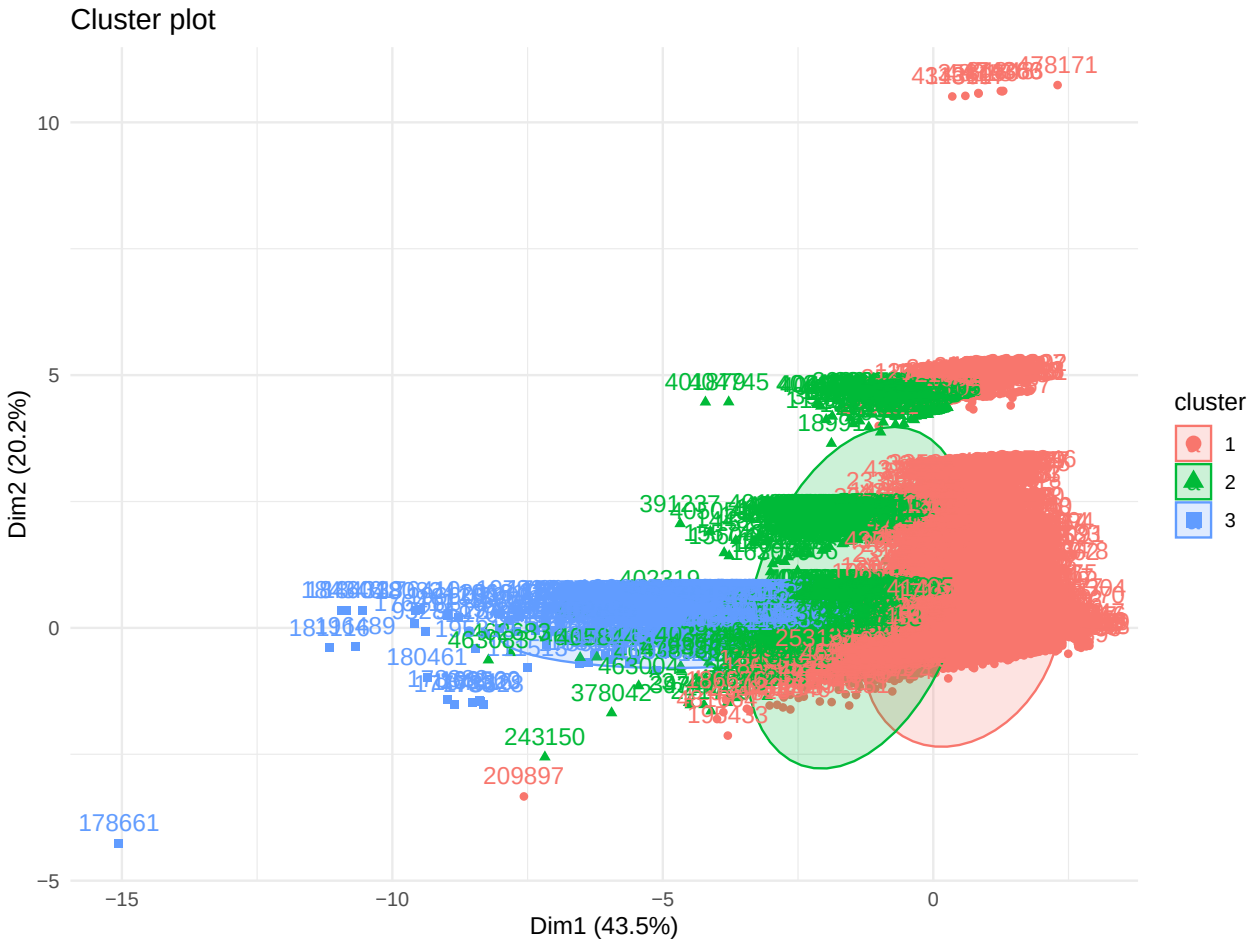
Algorithm steps:

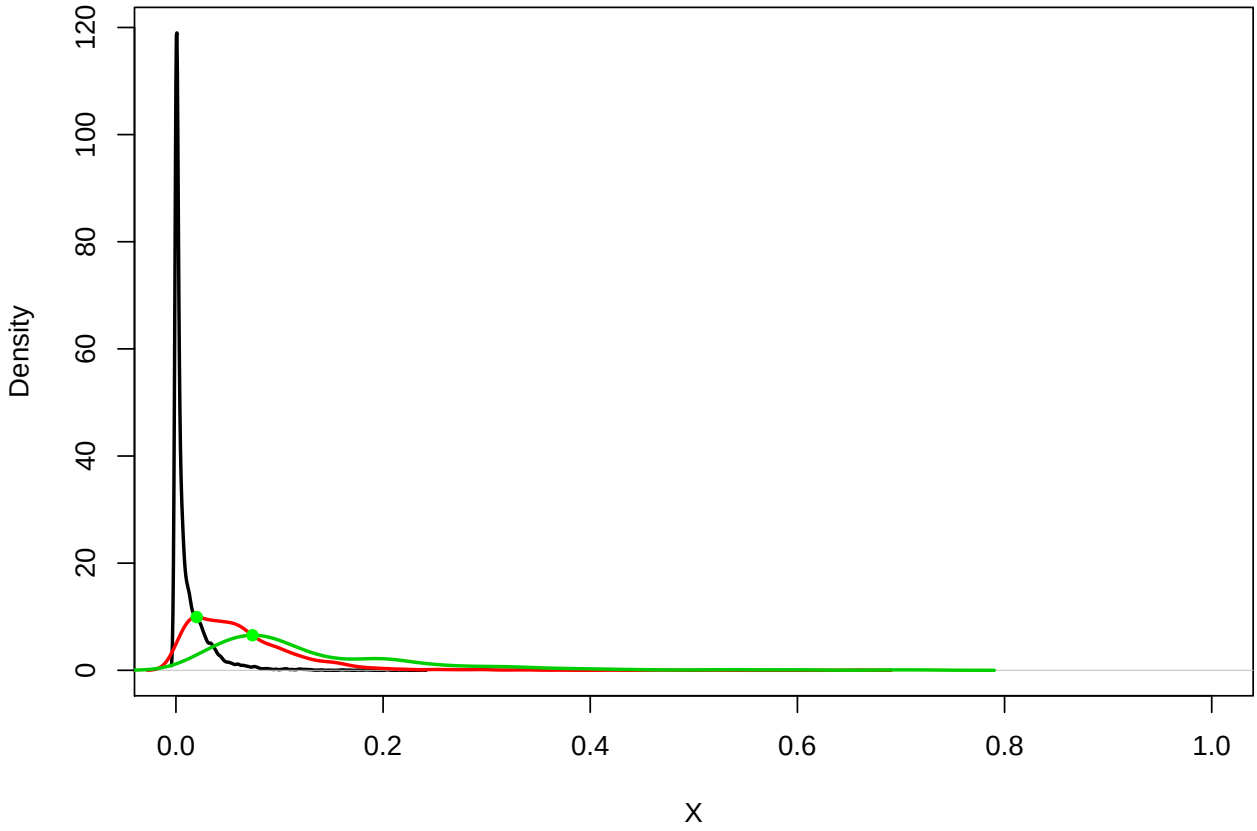
1. It replaces all NAs by random imputation, i.e., for each variable with missings, it draws from the marginal distribution of this variable not taking into account any correlations with other variables
2. Weights < 1 are used to adjust the scale of an observation that was generated in step 1. The weights are calculated by a (linear) weight function that starts near zero and converges to 1 at n_end.
3. A k-means clustering is performed with a number of c_steps steps starting with a random initialization.
4. The values from step 2 are replaced by new draws conditionally on the assigned cluster from step 3.
5. Steps 2-4 are repeated nr_iter times in total. The k-means clustering in step 3 uses the previous cluster centroids for initialization.
6. After the last draws a final k-means clustering is performed.

Amenities

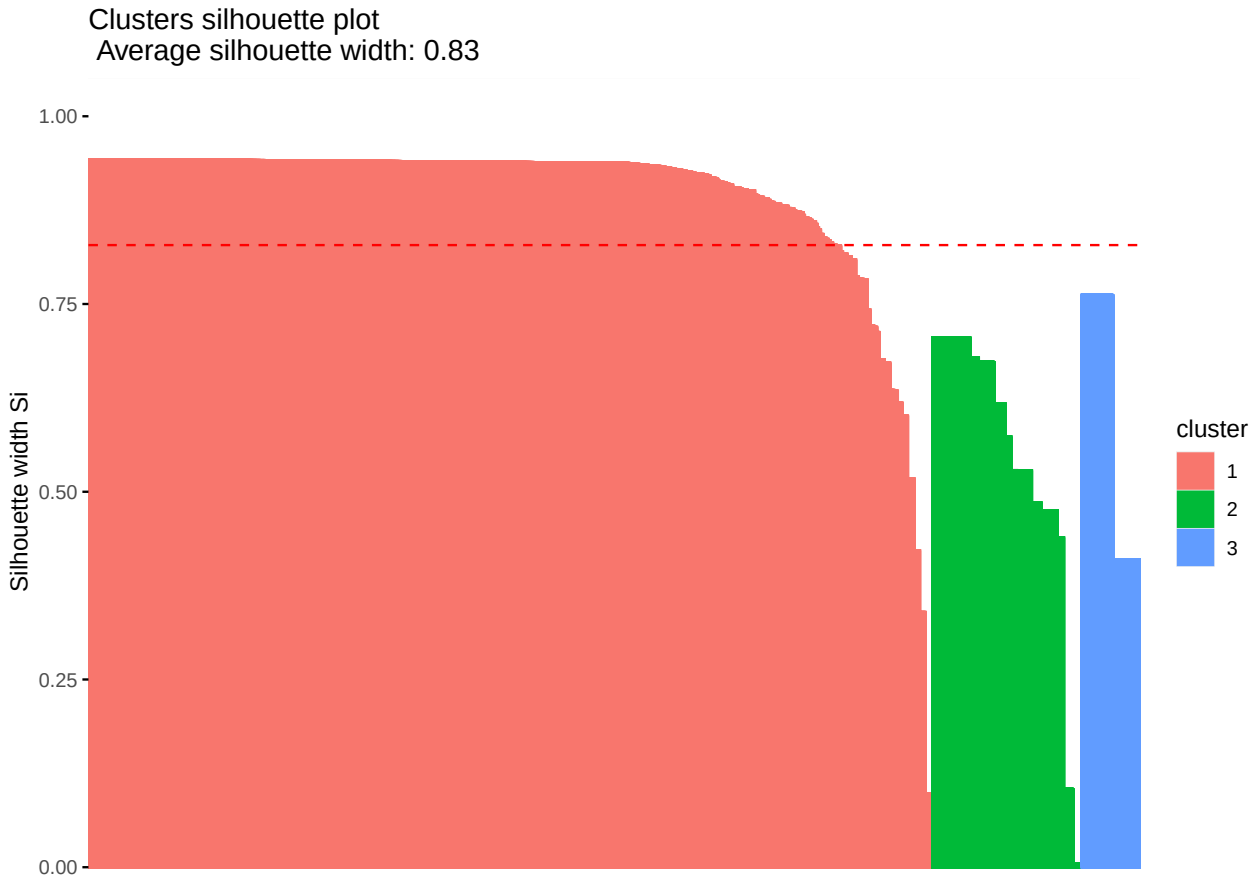
Employment




Cluster plot
Dim2 (20.2%)
Dim1 (43.5%)
cluster
1
2
3
10
5
0
−5
−15
−10
−5
0
178661
209897
243150
180461
378042



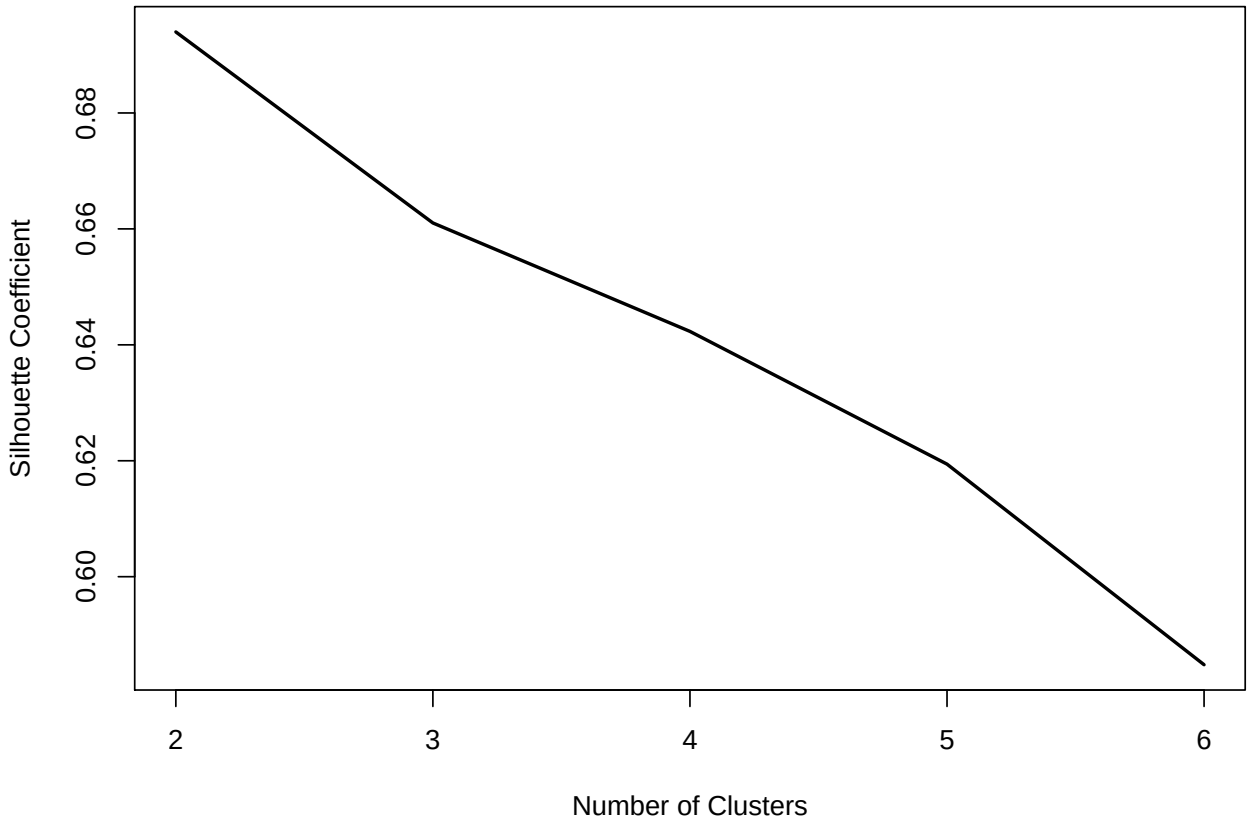

```
## [1] "Segment cutoff values:"
## [1] 0.02006458
## [1] 0.07388893
##   cluster  size ave.sil.width
## 1       1 10152          0.90
## 2       2  1798          0.54
## 3       3   703          0.60
```

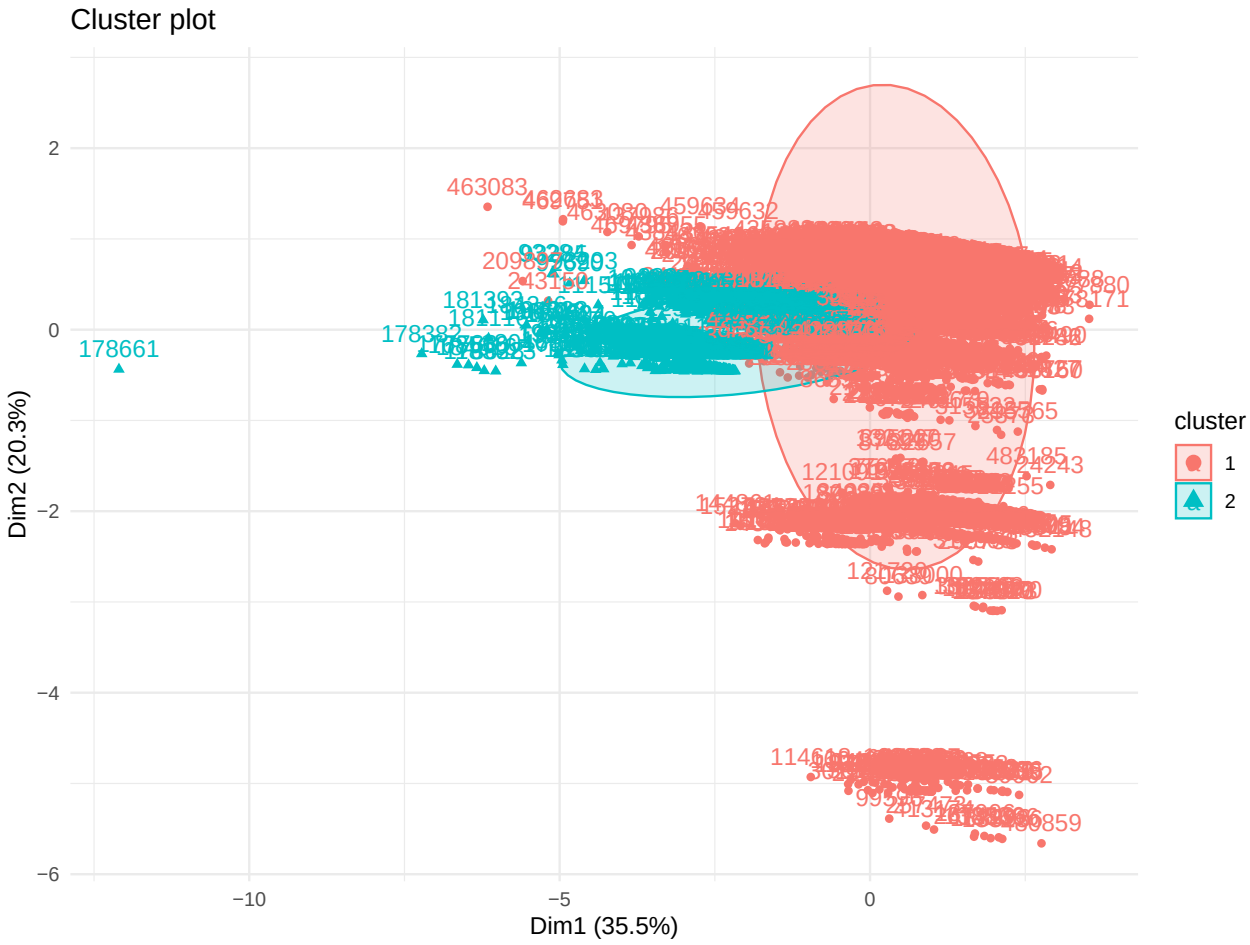


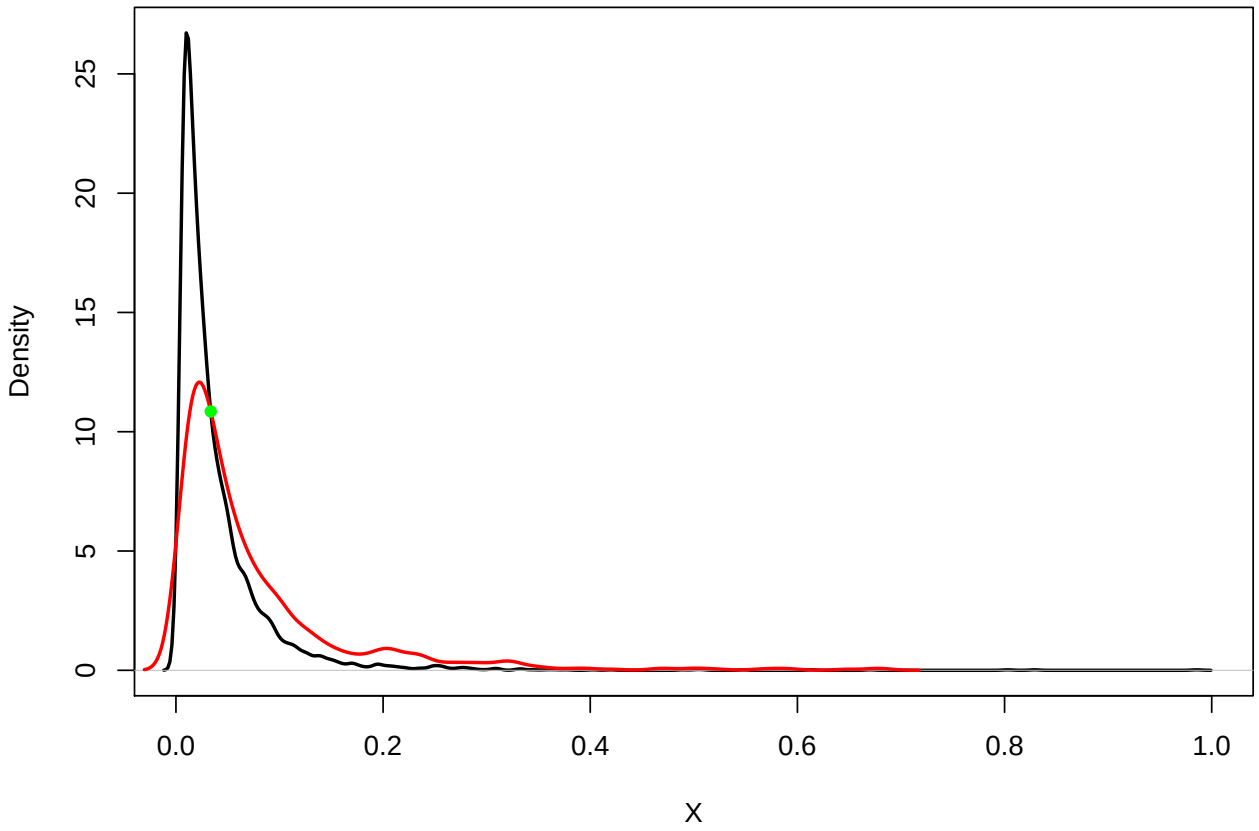

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2 Cluster 3
##      11790      2078       822
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2 Cluster 3
##         72      75.1      73.5
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2 Cluster 3
##   236132.1  217224.7  251392.2
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2 Cluster 3
##        5176       914       348
## B      4897       868       350
## D      1302       222        94
## K       415        74        30
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2 Cluster 3
##      0.228     0.231      0.23
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2 Cluster 3
## Alberta                     368        66        24
## BritishColumbia             581       115        27
## NewBrunswick                 93        20         9
## NorthwestTerritories          7         0         0
## NovaScotia                  363        59        28
## Ontario                    1805       302       133
## Quebec                      694       113        51
## Saskatchewan                 61        10         0
## NA's                       7818      1393       550
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2 Cluster 3
## 0     10669      1887       738
## 1       856       160        68
## 2       133         8         6
## F       132        23        10
```

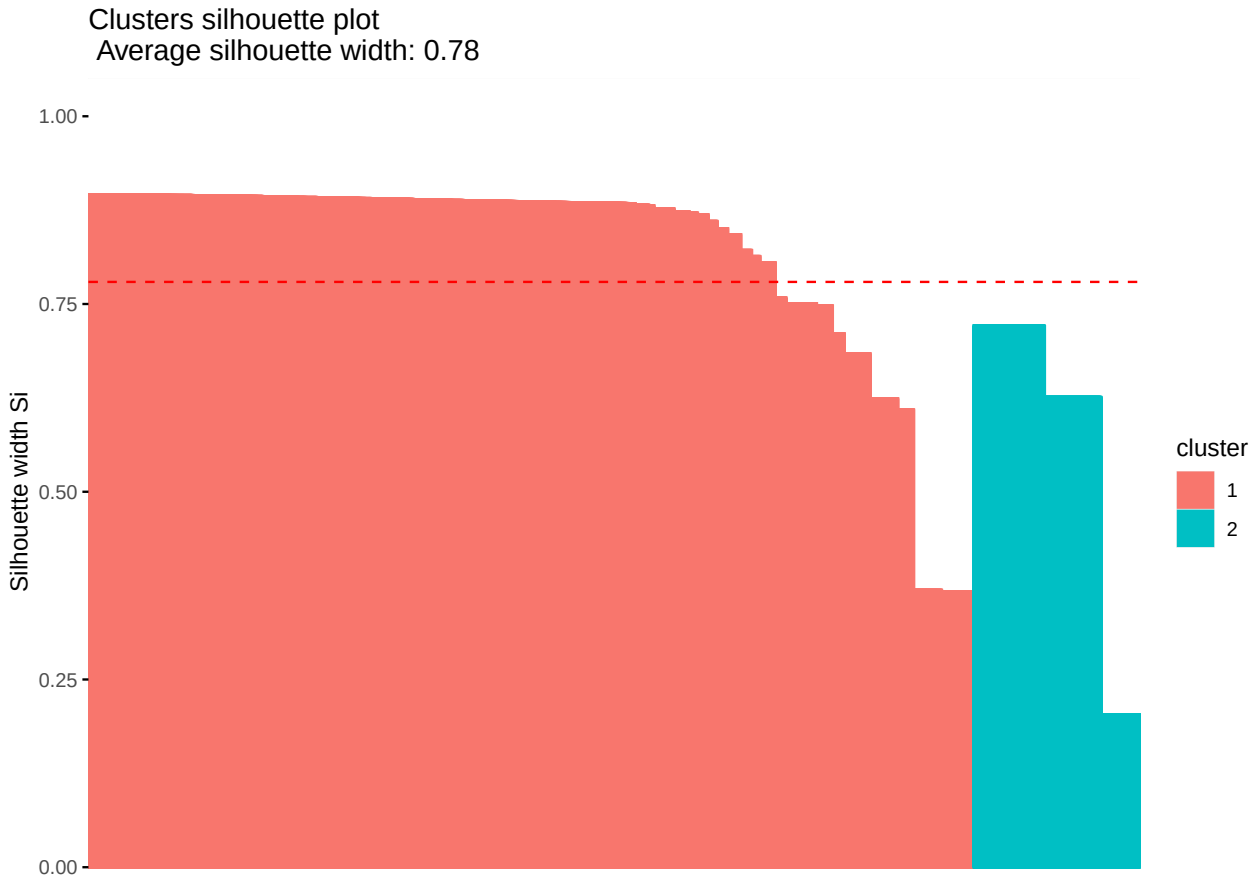
Pharmacy


Silhouette Coefficient
0.60
0.62
0.64
0.66
0.68
2
3
4
5
6
Number of Clusters


Cluster plot
Dim2 (20.3%)
Dim1 (35.5%)
2
0
−2
−4
−6
−10
−5
0
cluster
1
2
178661
463083



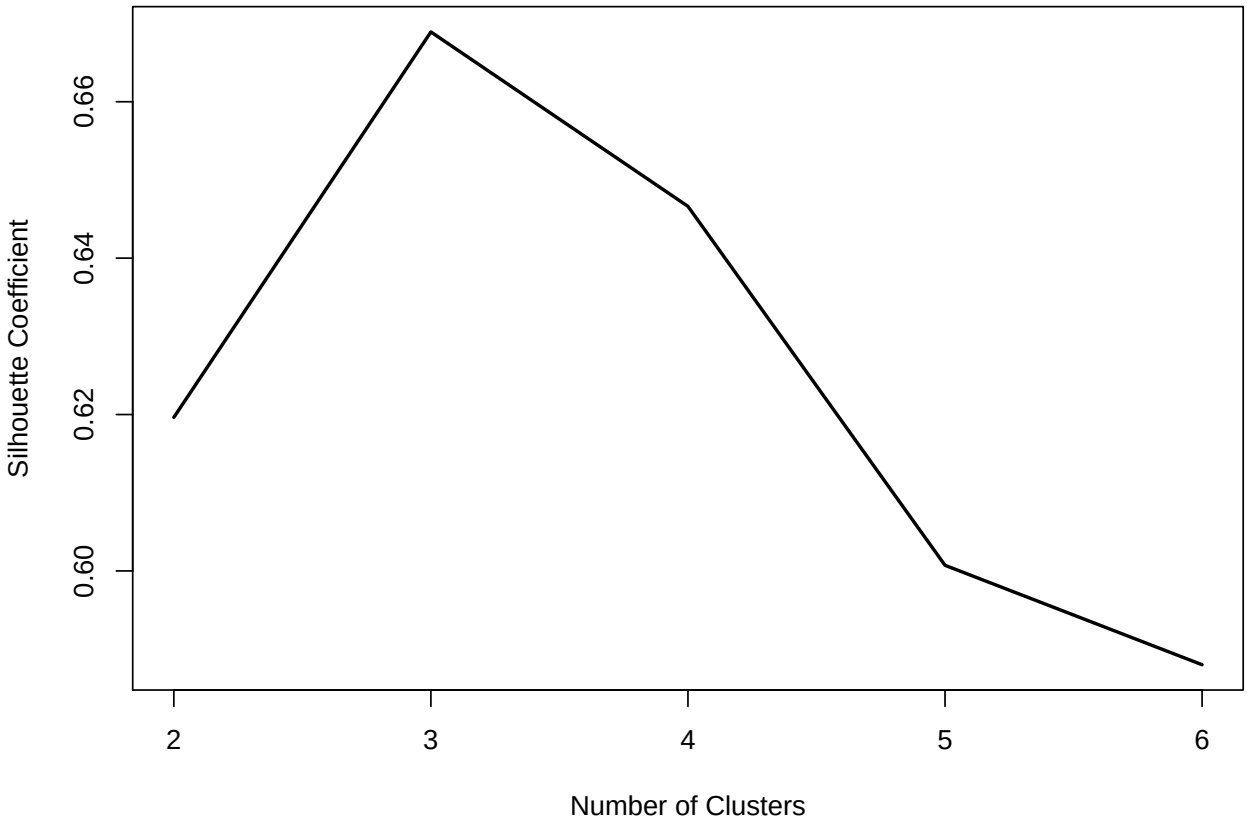

```
## [1] "Segment cutoff values:"
## [1] 0.03363201
##   cluster size ave.sil.width
## 1       1 4403          0.82
## 2       2  828          0.57
```

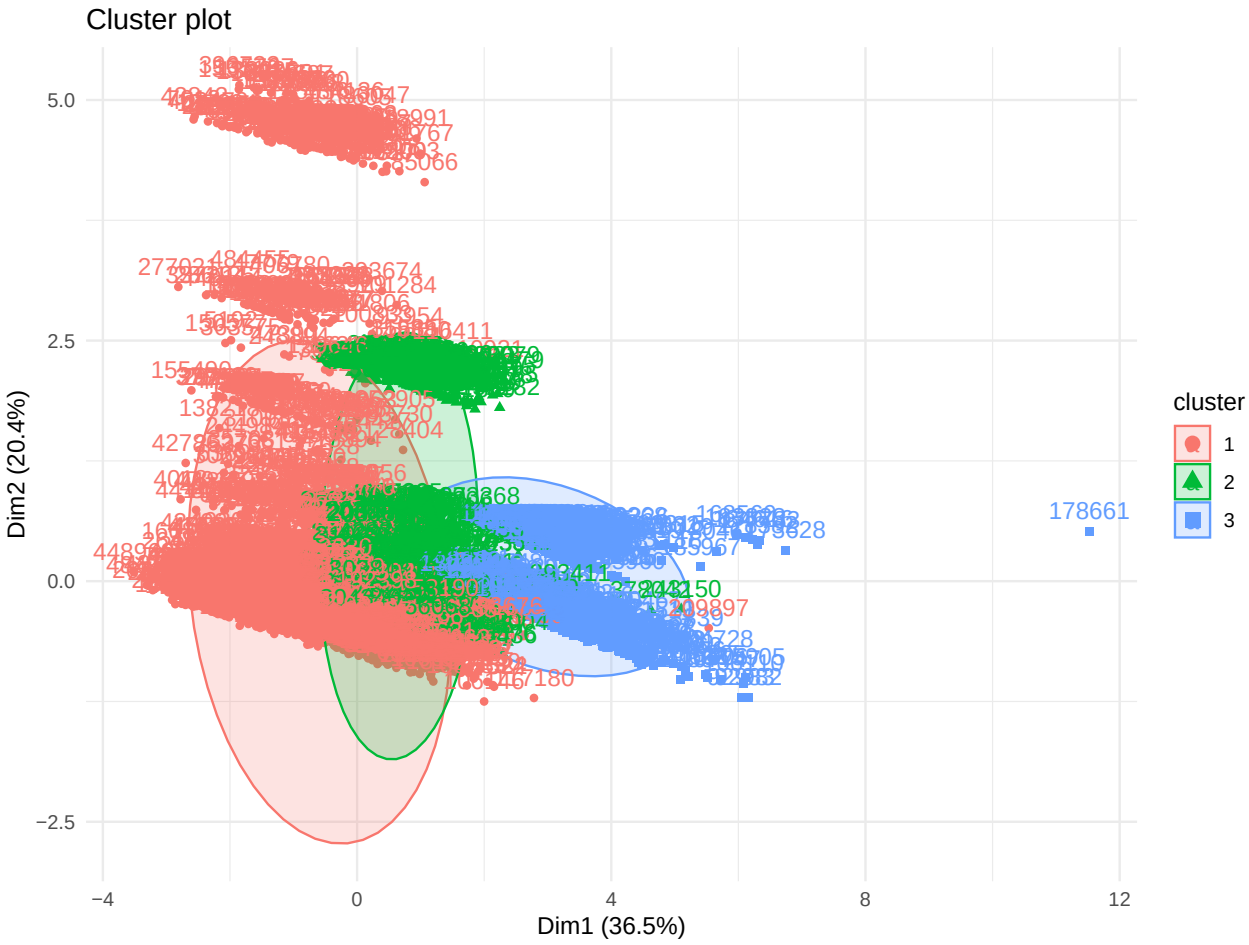


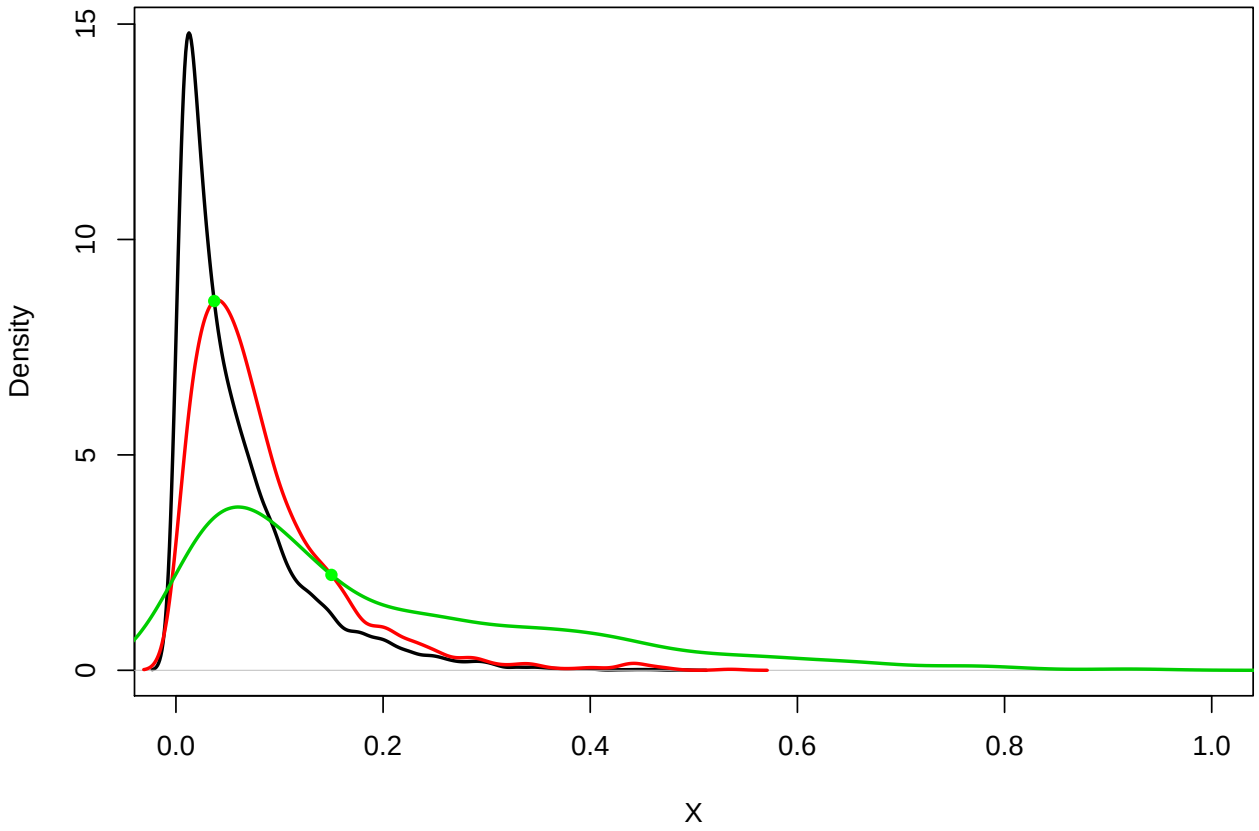

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##      12341      2349
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       73.7      66.3
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   233788.4  237056.2
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2
##        5422      1016
## B      5112      1003
## D      1361       257
## K       446        73
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2
##      0.229     0.226
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2
## Alberta                     375        83
## BritishColumbia             617       106
## NewBrunswick                 99        23
## NorthwestTerritories          6         1
## NovaScotia                  374        76
## Ontario                    1881       359
## Quebec                      727       131
## Saskatchewan                 62         9
## NA's                       8200      1561
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0     11175      2119
## 1       900       184
## 2       123        24
## F       143        22
```

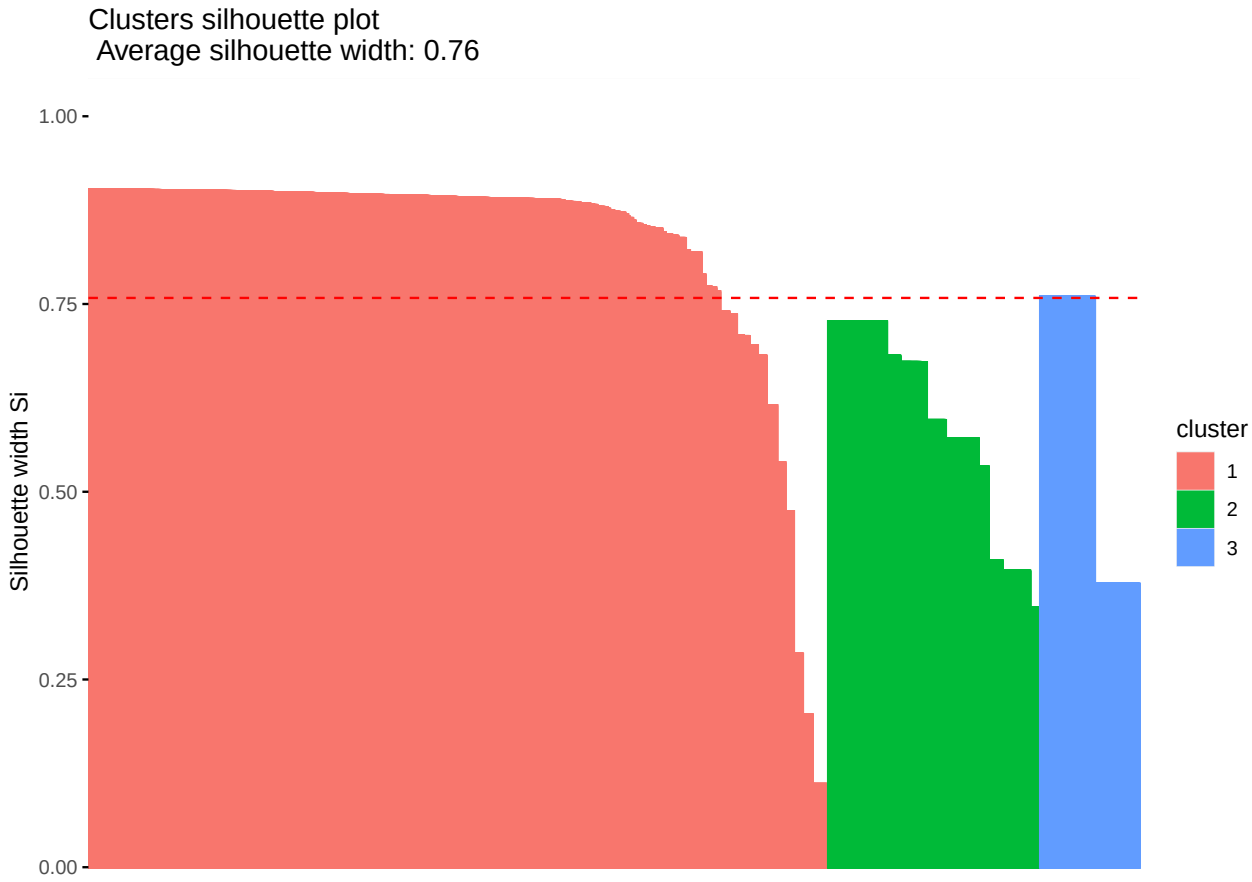
Childcare


Silhouette Coefficient
0.60
0.62
0.64
0.66
2
3
4
5
6
Number of Clusters


Cluster plot
Dim2 (20.4%)
Dim1 (36.5%)
5.0
2.5
0.0
−2.5
−4
0
4
8
12
cluster
1
2
3
178661



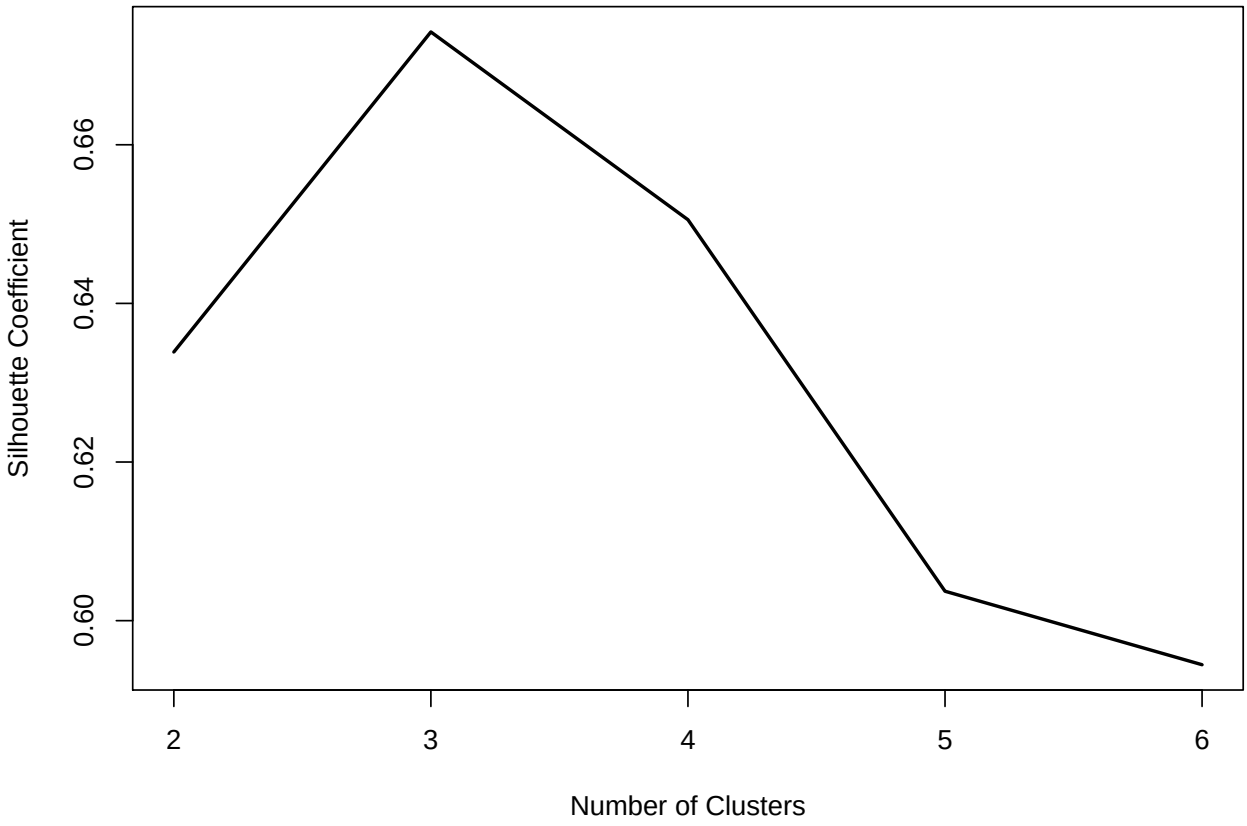

```
## [1] "Segment cutoff values:"
## [1] 0.03697689
## [1] 0.1501263
##   cluster size ave.sil.width
## 1       1 5086          0.83
## 2       2 1458          0.59
## 3       3  687          0.59
```

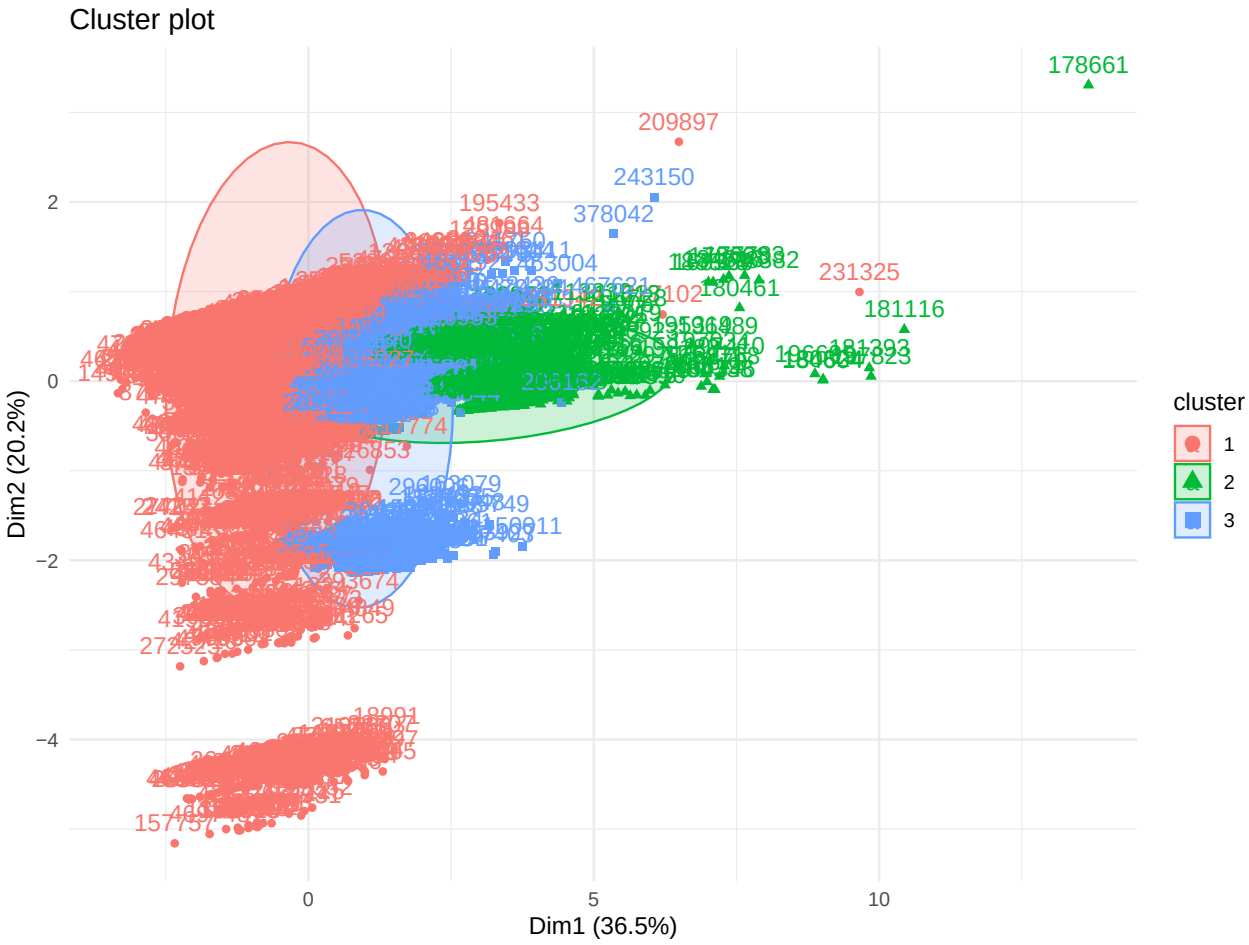


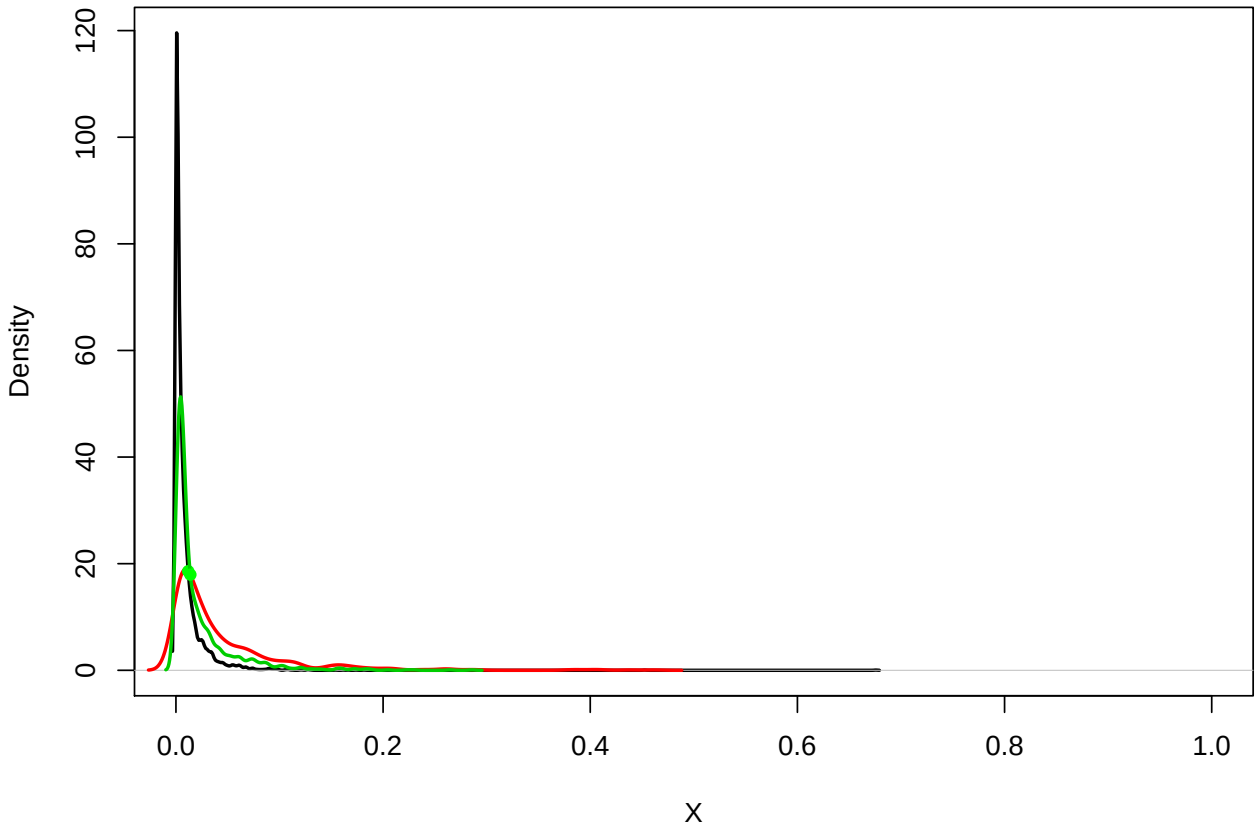

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2 Cluster 3
##      10332      2962      1396
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2 Cluster 3
##       71.8      76.1        70
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2 Cluster 3
##   233477.7    233613  241953.1
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2 Cluster 3
##        4572      1255       611
## B      4285      1246       584
## D      1118       351       149
## K       357       110        52
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2 Cluster 3
##      0.228     0.229      0.23
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2 Cluster 3
## Alberta                     334        82        42
## BritishColumbia             513       148        62
## NewBrunswick                 89        21        12
## NorthwestTerritories          6         1         0
## NovaScotia                  322        95        33
## Ontario                    1572       450       218
## Quebec                      602       193        63
## Saskatchewan                 48        16         7
## NA's                       6846      1956       959
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2 Cluster 3
## 0      9359      2683      1252
## 1       757       214       113
## 2        97        39        11
## F       119        26        20
```

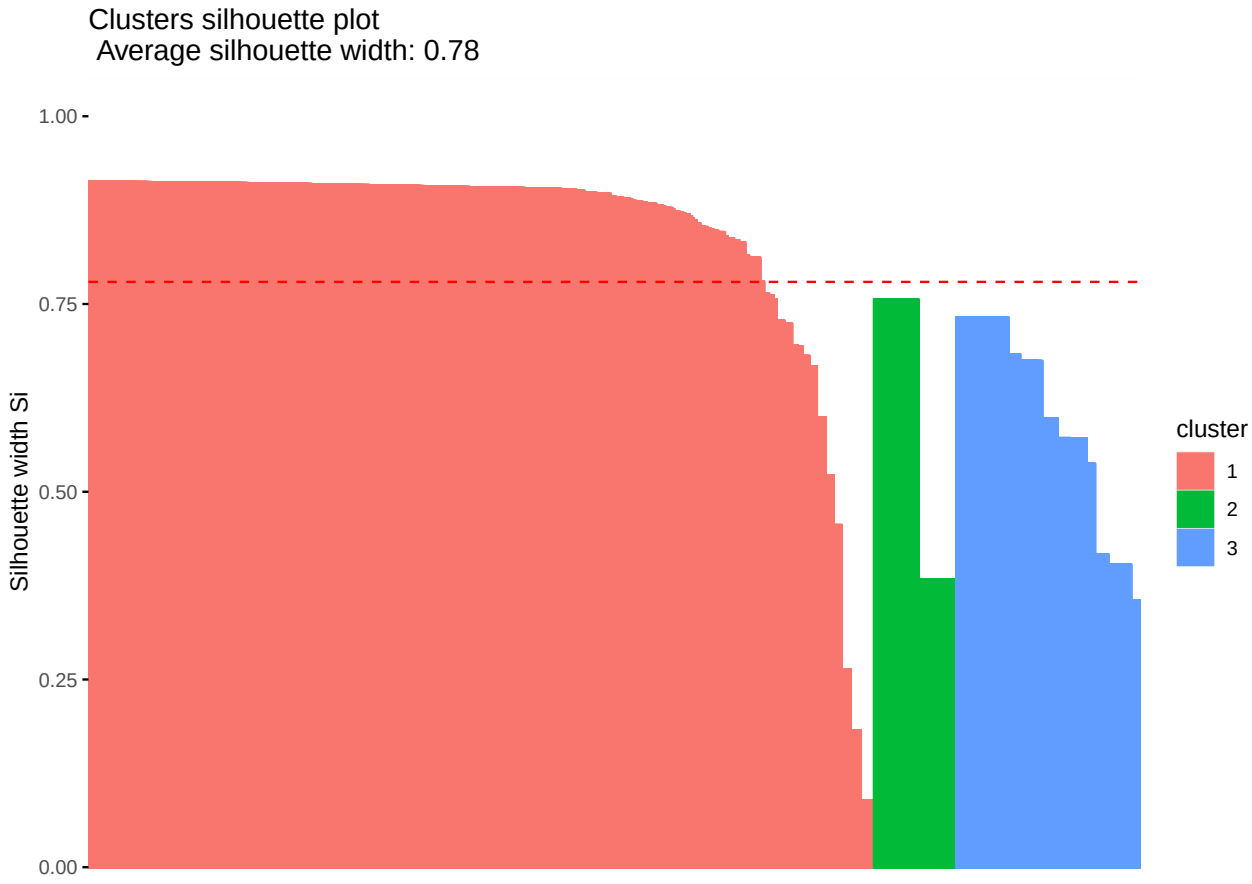
Health


Silhouette Coefficient
0.60
0.62
0.64
0.66
2
3
4
5
6
Number of Clusters


Cluster plot
Dim2 (20.2%)
Dim1 (36.5%)
cluster
1
2
3
178661
209897
243150
195433
378042
231325
180461
181116
181393
203162
0
5
10
2
0
−2
−4



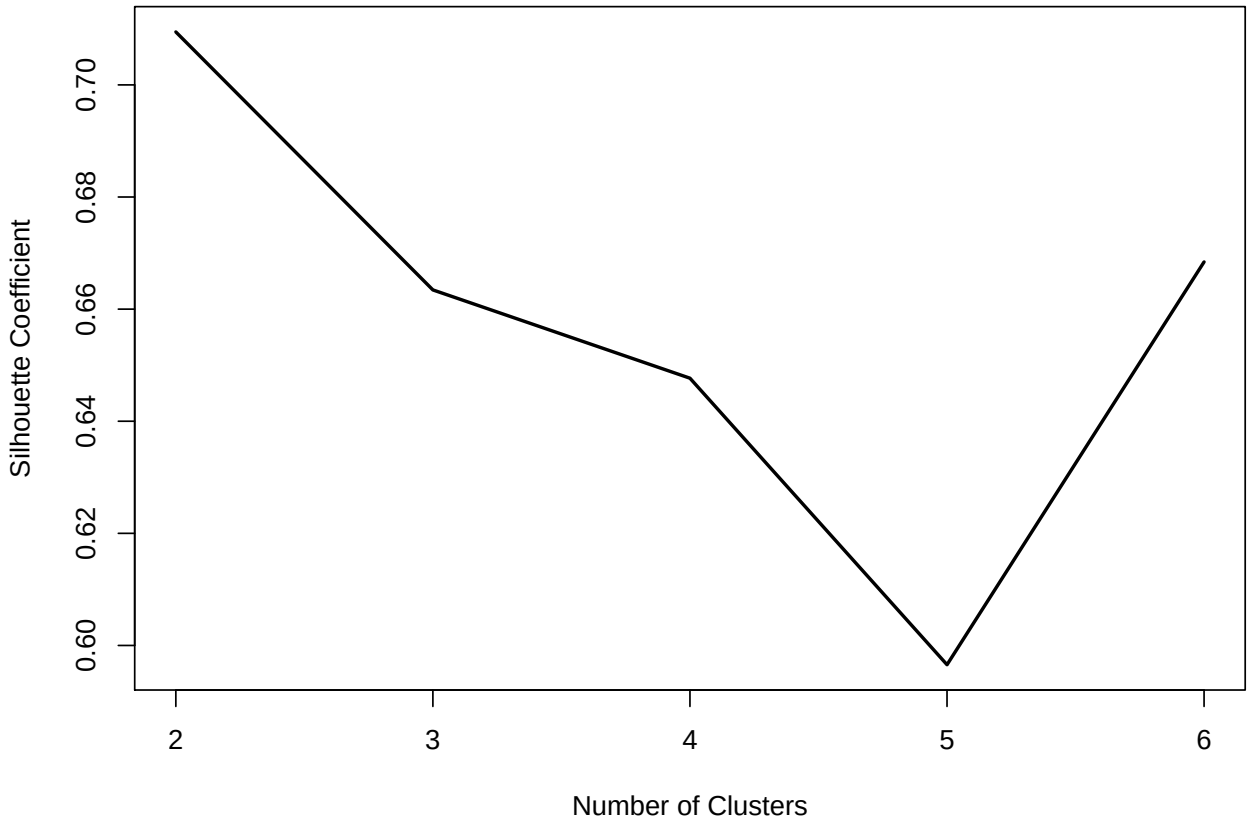

```
## [1] "Segment cutoff values:"
## [1] 0.01186657
## [1] 0.01393482
##   cluster size ave.sil.width
## 1       1 6658          0.84
## 2       2  697          0.59
## 3       3 1558          0.59
```



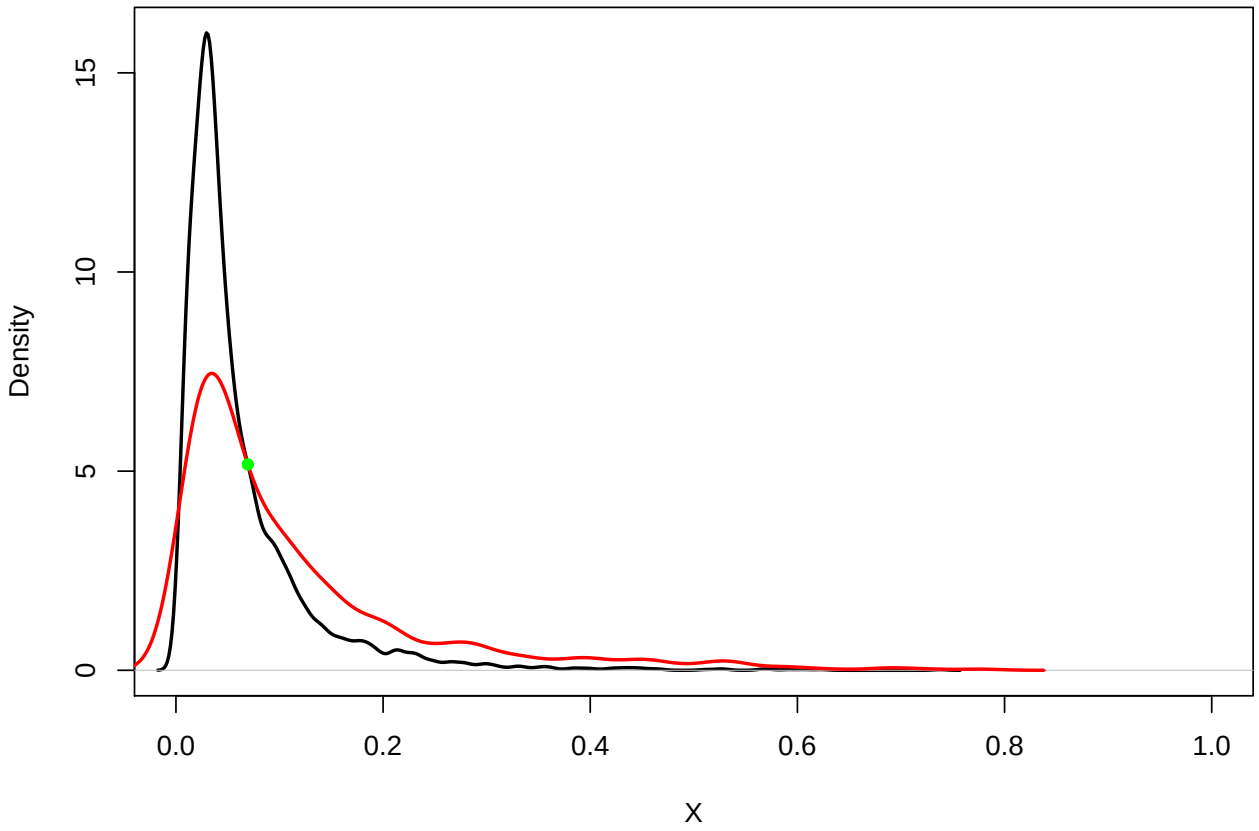

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2 Cluster 3
##      10956      1175      2559
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2 Cluster 3
##       72.2      79.7      70.8
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2 Cluster 3
##   236926.1  228226.7  225904.2
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2 Cluster 3
##        4815       546      1077
## B      4563       465      1087
## D      1199       123       296
## K       379        41        99
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2 Cluster 3
##      0.228     0.236     0.228
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2 Cluster 3
## Alberta                     319        29       110
## BritishColumbia             530        59       134
## NewBrunswick                100         4        18
## NorthwestTerritories          5         0         2
## NovaScotia                  337        32        81
## Ontario                    1671       176       393
## Quebec                      644        71       143
## Saskatchewan                 53         9         9
## NA's                       7297       795      1669
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2 Cluster 3
## 0      9927      1052      2315
## 1       793        99       192
## 2       111        12        24
## F       125        12        28
```

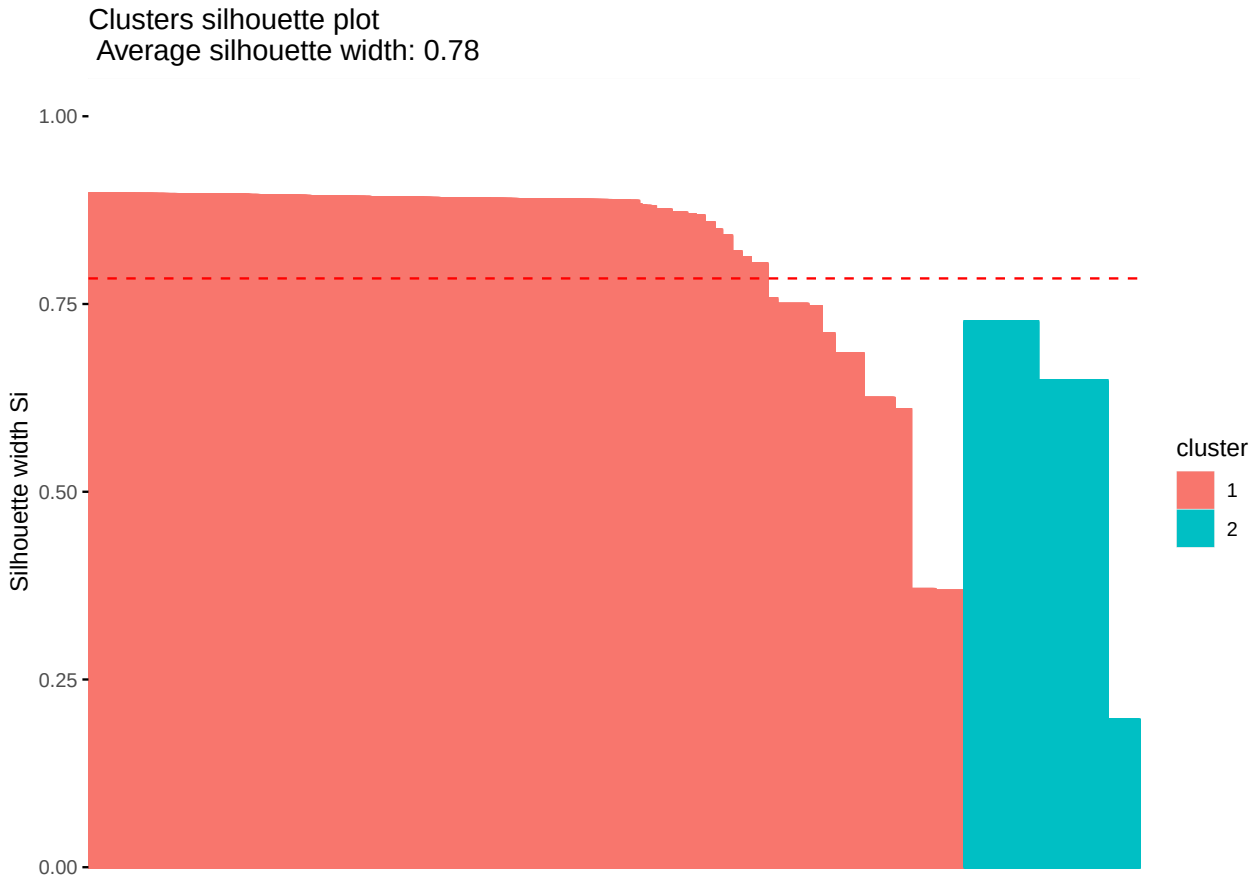
Grocery


Silhouette Coefficient
0.60
0.62
0.64
0.66
0.68
0.70
2
3
4
5
6
Number of Clusters


Cluster plot
Dim2 (20.2%)
Dim1 (36.3%)
cluster
1
2
2
0
−2
−4
−6
−2.5
0.0
2.5
5.0
7.5
243150
378042
451151
79115
138000
452122
288339
18991



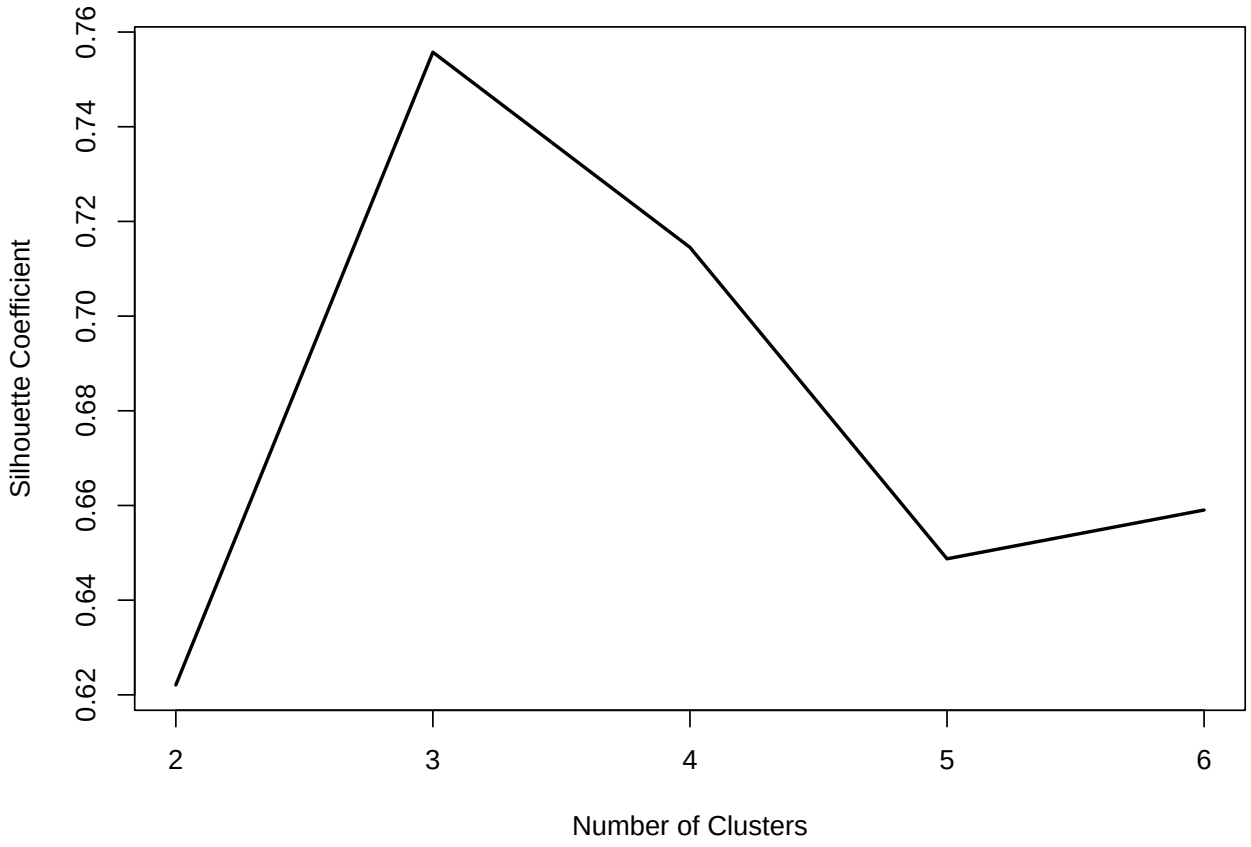

```
## [1] "Segment cutoff values:"
## [1] 0.06941786
##   cluster size ave.sil.width
## 1       1 3495          0.82
## 2       2  700          0.60
```

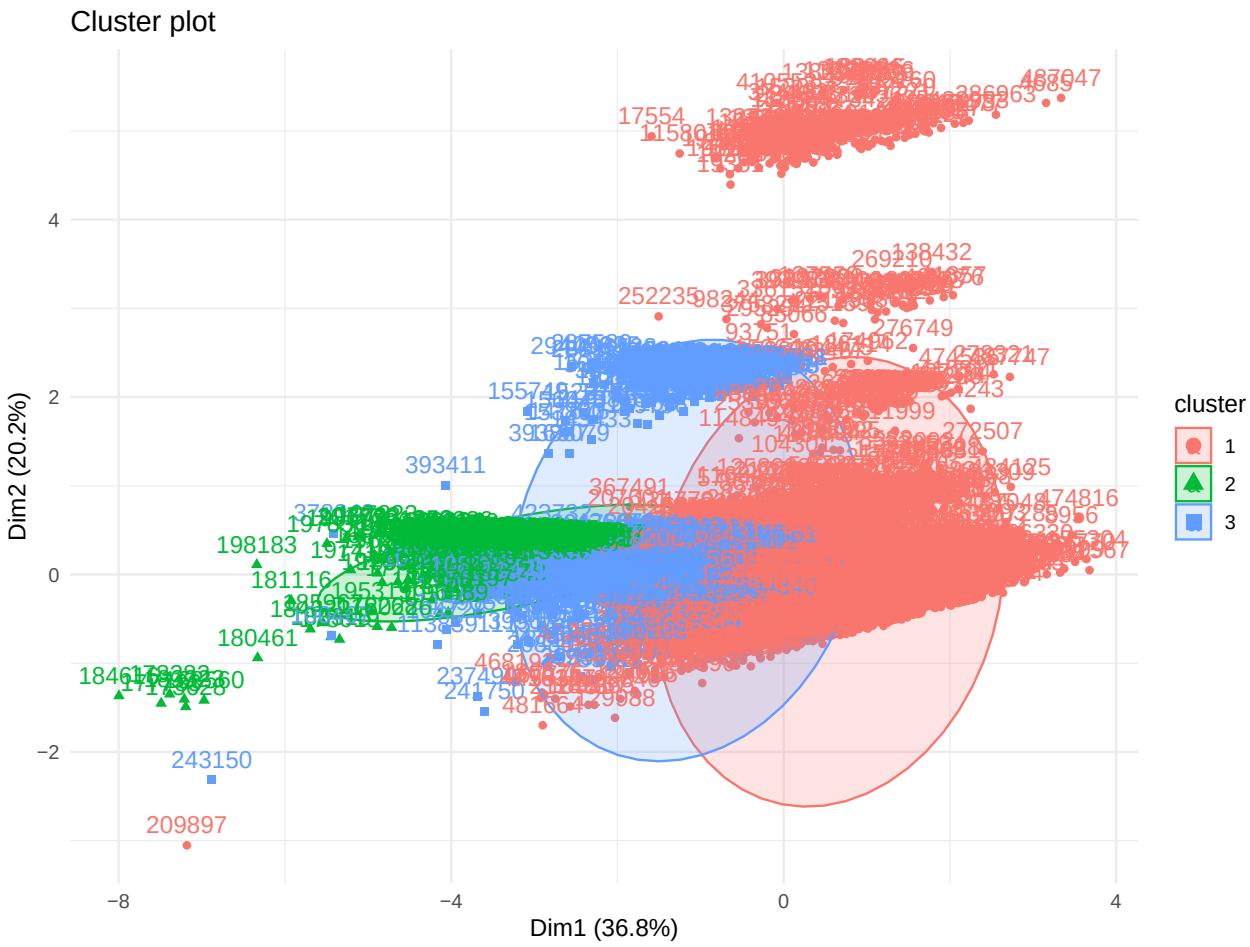


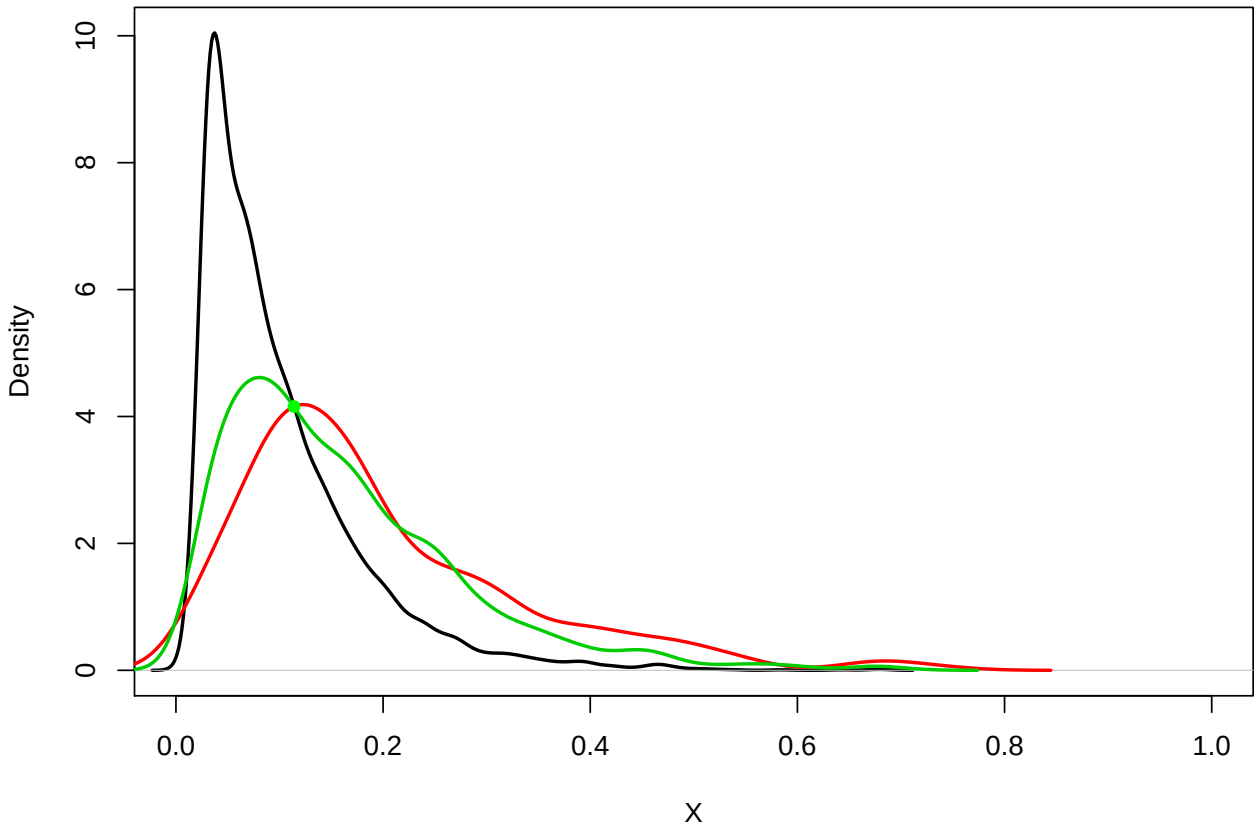

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##      12218      2472
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       72.5      72.6
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   232517.5  243173.9
##
##
##
##  CMA Type:
##    Cluster 1 Cluster 2
##         5378      1060
## B       5062      1053
## D       1346       272
## K        432        87
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2
##      0.229     0.224
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2
## Alberta                     377        81
## BritishColumbia             588       135
## NewBrunswick                102        20
## NorthwestTerritories          6         1
## NovaScotia                  372        78
## Ontario                    1871       369
## Quebec                      706       152
## Saskatchewan                 52        19
## NA's                       8144      1617
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0     11076      2218
## 1       889       195
## 2       118        29
## F       135        30
```

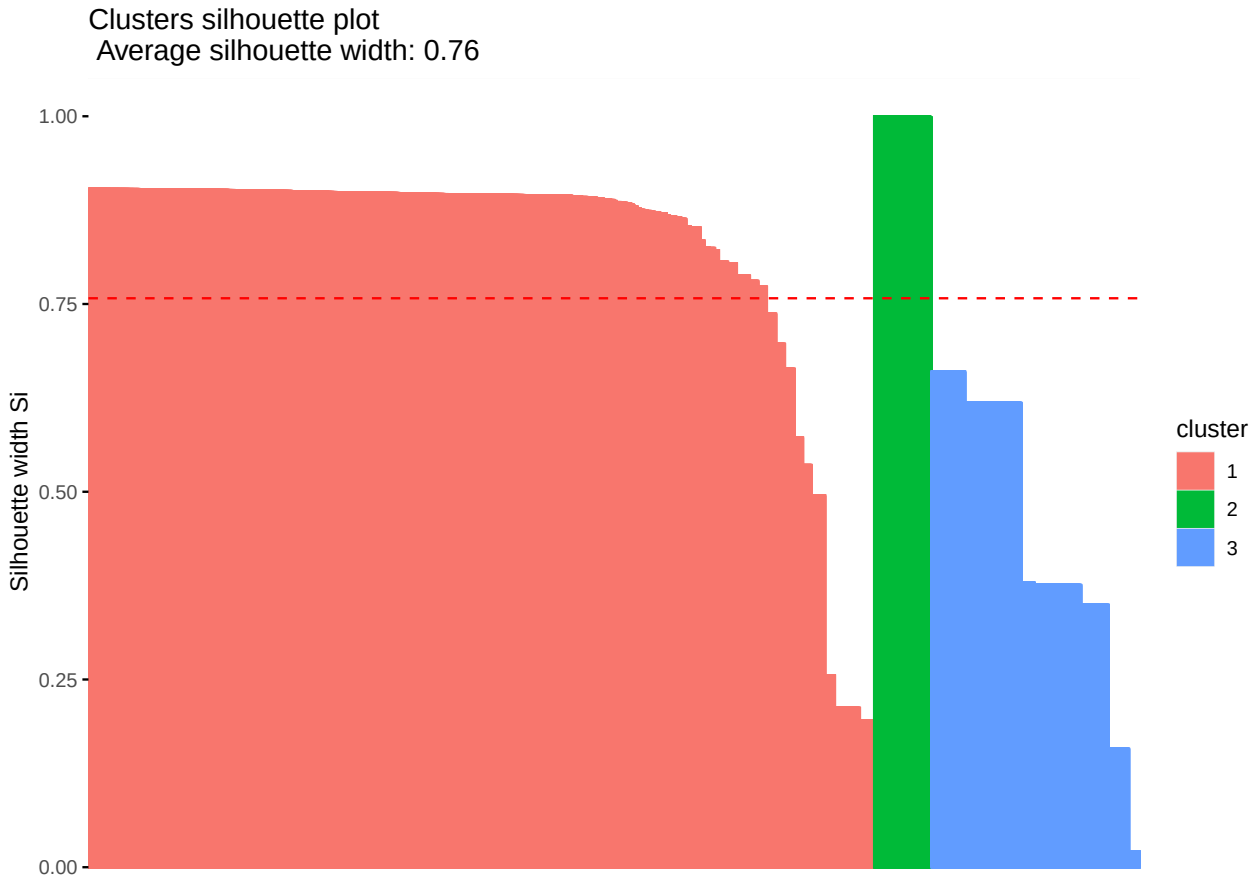
Primary Education




Cluster plot
Dim2 (20.2%)
Dim1 (36.8%)
cluster
1
2
3
4
2
0
−2
−8
−4
0
4
393411
198183
181116
180461
243150
209897



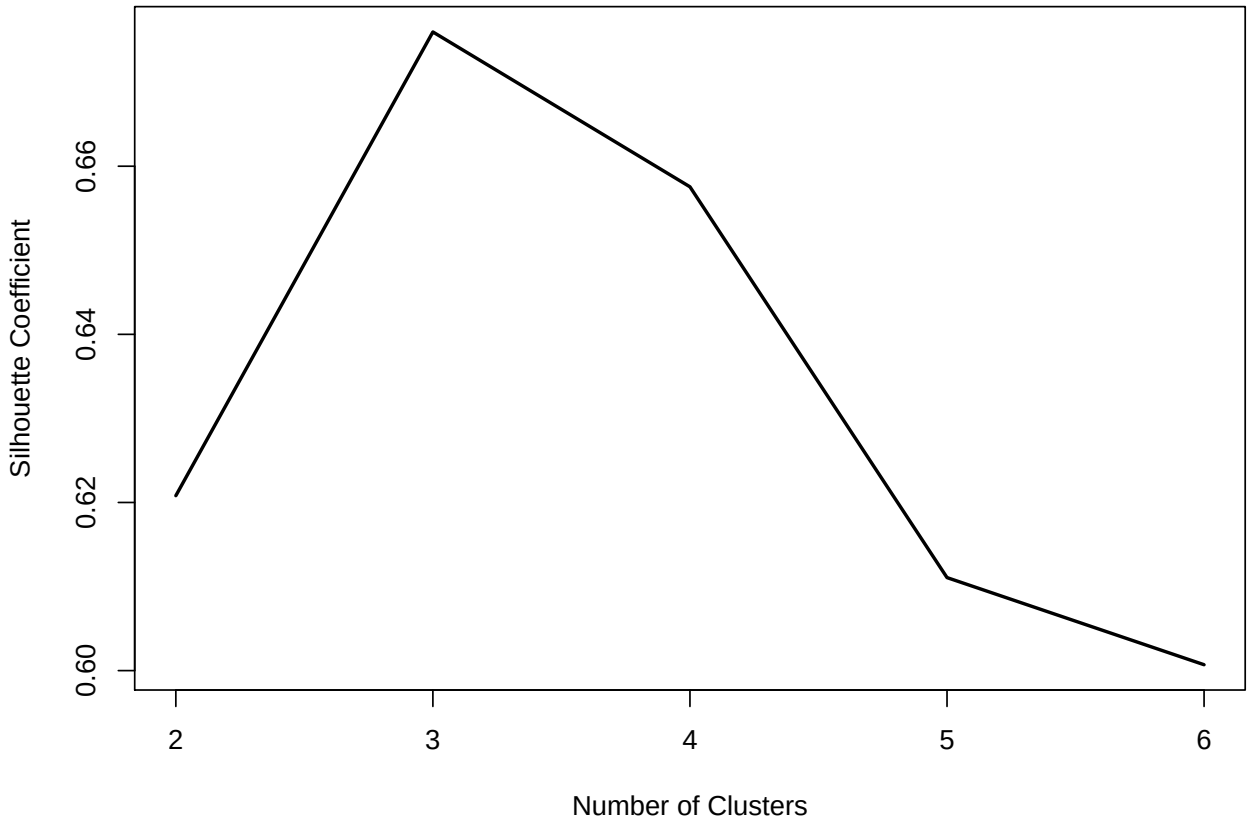

```
## [1] "Segment cutoff values:"
## [1] 0.1141726
## [1] 0.11379
##   cluster size ave.sil.width
## 1       1 4999          0.82
## 2       2  365          1.00
## 3       3 1327          0.44
```

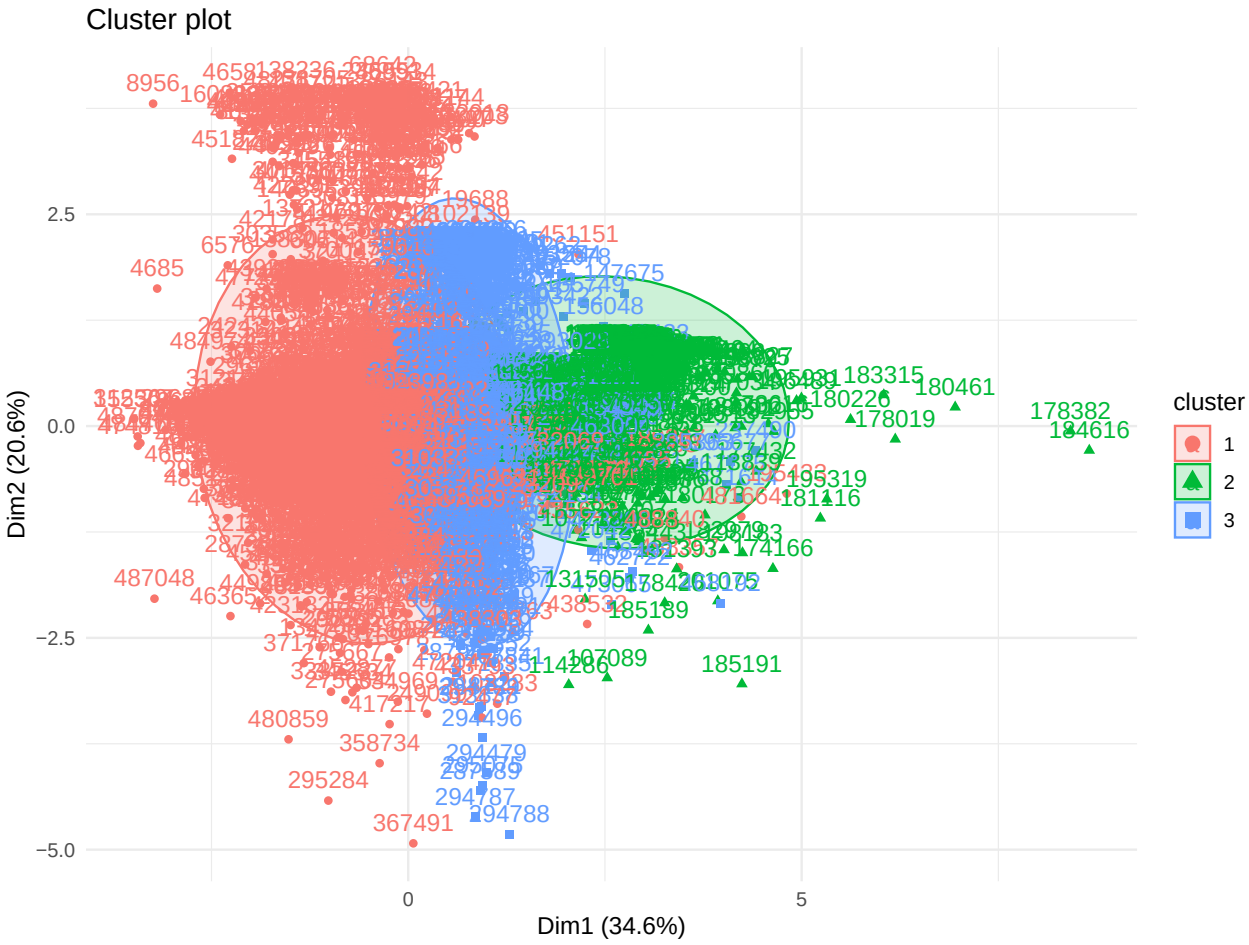


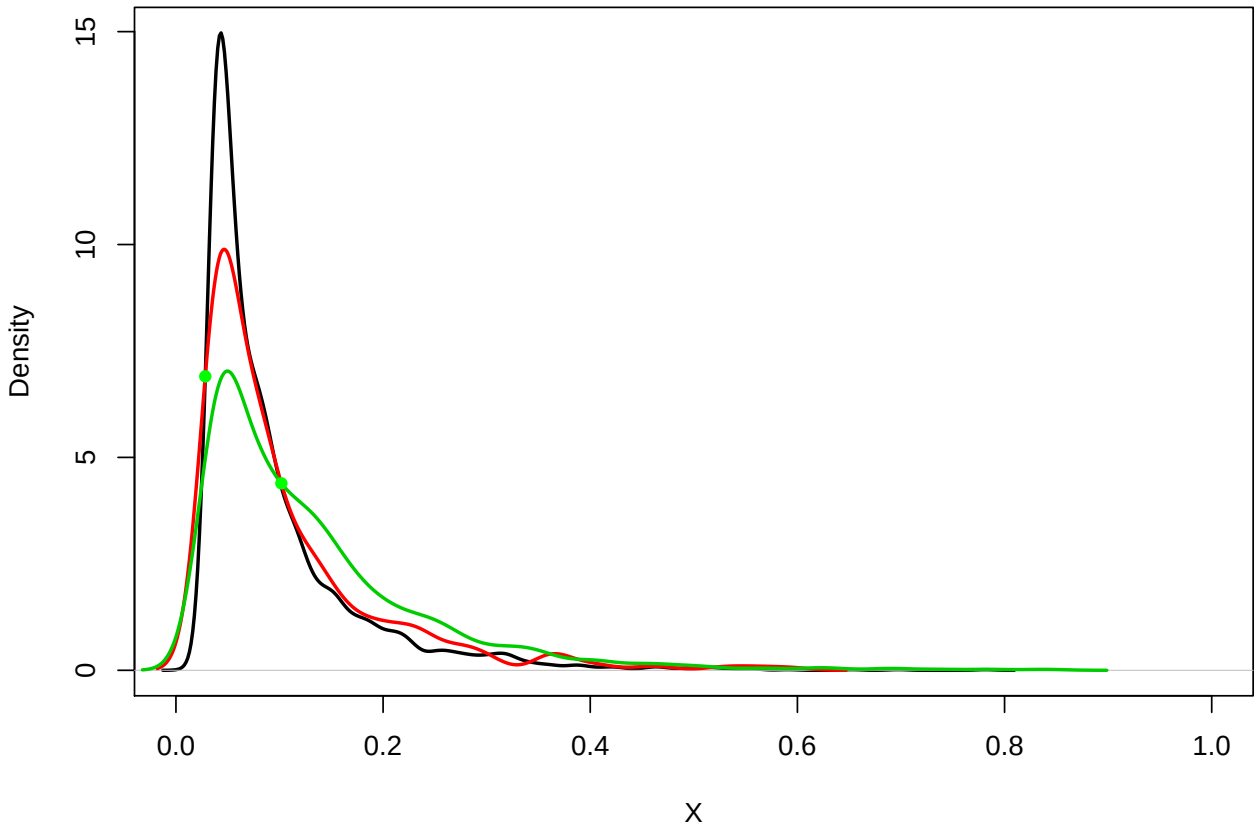

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2 Cluster 3
##      10975       815      2900
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2 Cluster 3
##       73.1      69.7      71.2
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2 Cluster 3
##   237151.1  205277.9  231726.4
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2 Cluster 3
##        4797       359      1282
## B      4586       328      1201
## D      1197       106       315
## K       395        22       102
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2 Cluster 3
##      0.228     0.233     0.228
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2 Cluster 3
## Alberta                     337        27        94
## BritishColumbia             547        45       131
## NewBrunswick                 96         8        18
## NorthwestTerritories          4         1         2
## NovaScotia                  341        25        84
## Ontario                    1667       127       446
## Quebec                      652        42       164
## Saskatchewan                 52         5        14
## NA's                       7279       535      1947
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2 Cluster 3
## 0      9935       739      2620
## 1       808        60       216
## 2       112         4        31
## F       120        12        33
```

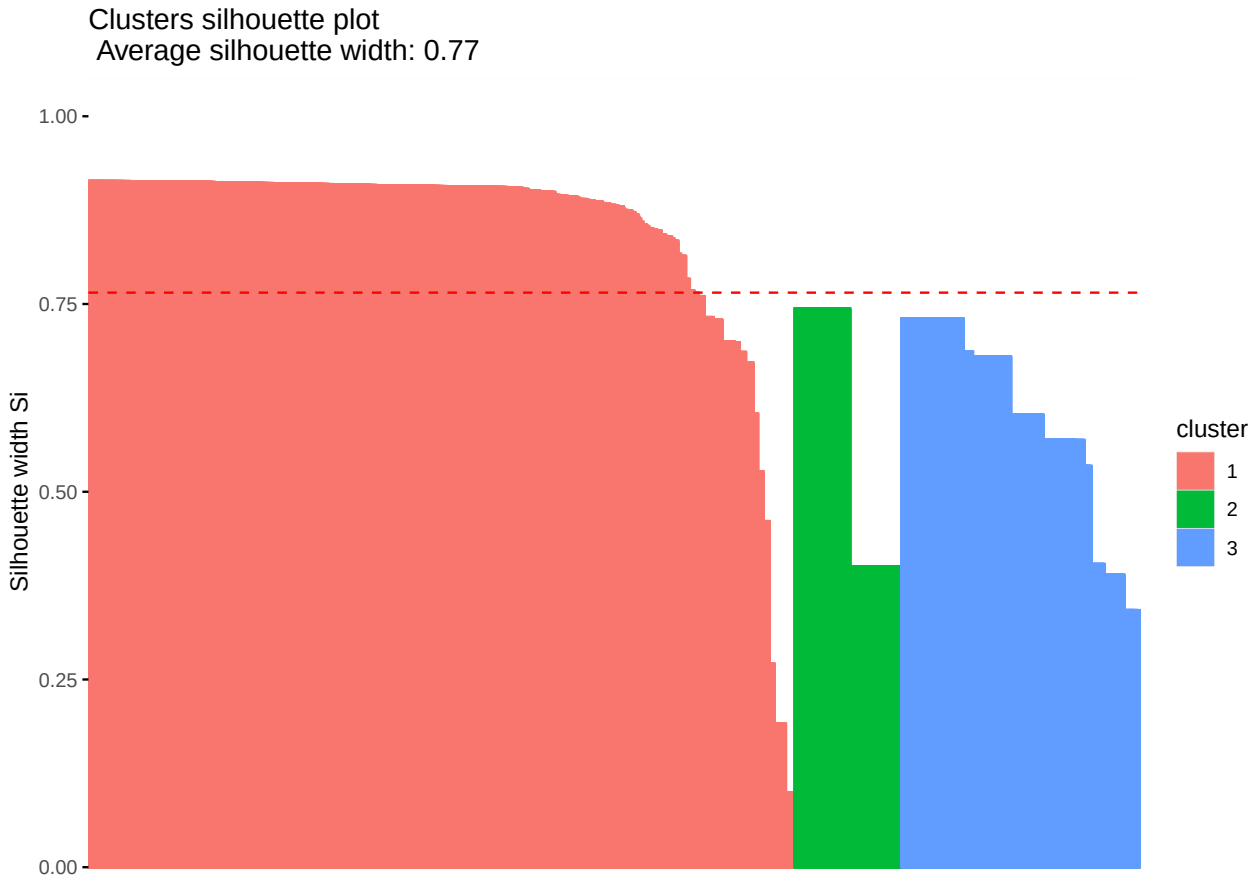
Secondary Education




Cluster plot
Dim2 (20.6%)
Dim1 (34.6%)
2.5
0.0
−2.5
−5.0
0
5
cluster
1
2
3



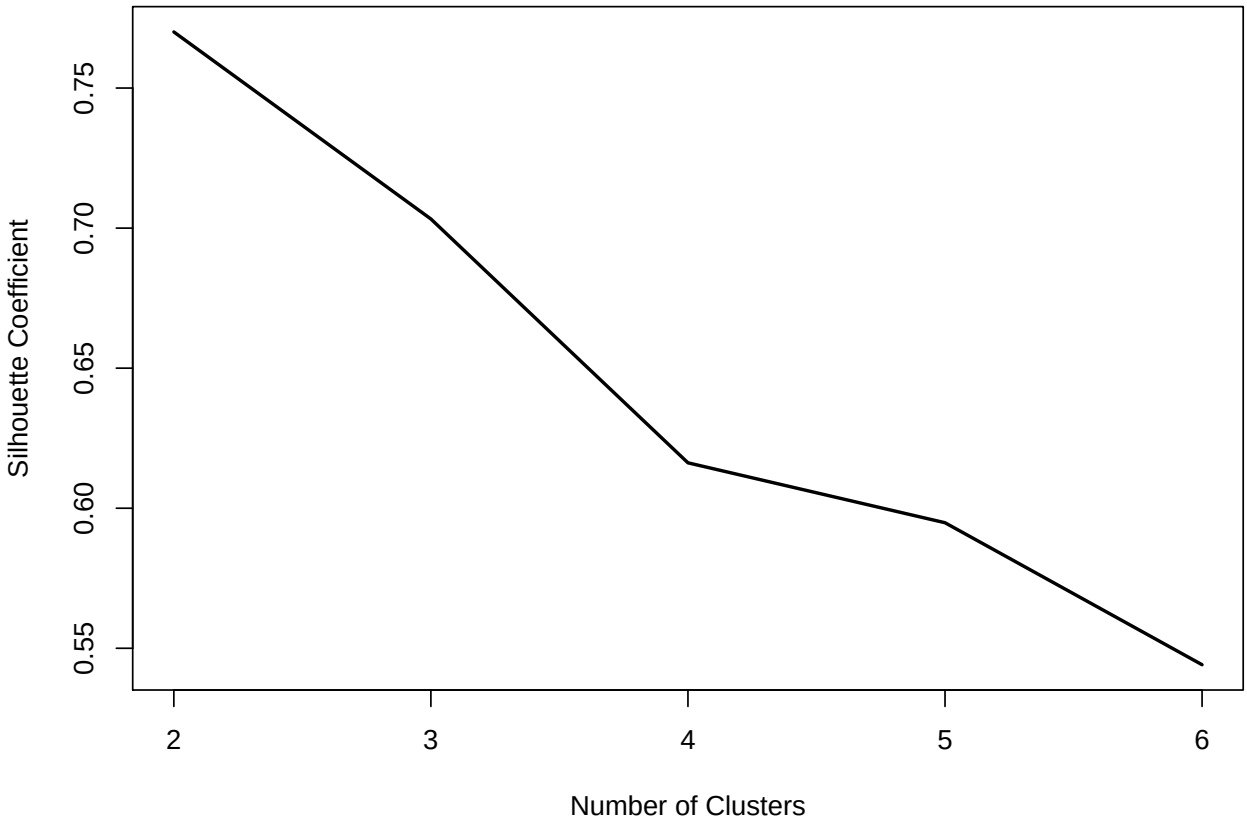

```
## [1] "Segment cutoff values:"
## [1] 0.02832299
## [1] 0.1017901
##   cluster size ave.sil.width
## 1       1 2813          0.85
## 2       2  425          0.58
## 3       3  951          0.60
```

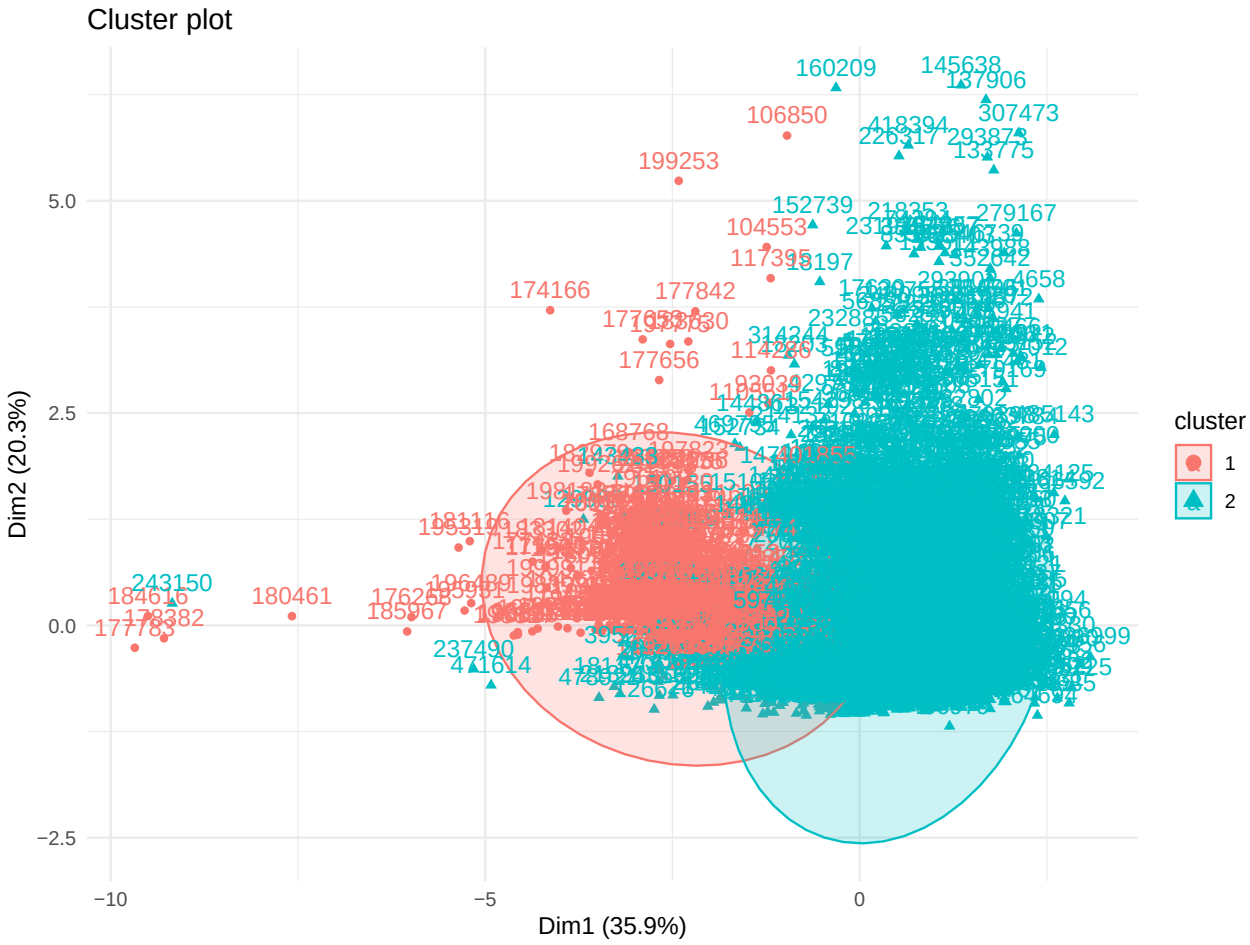


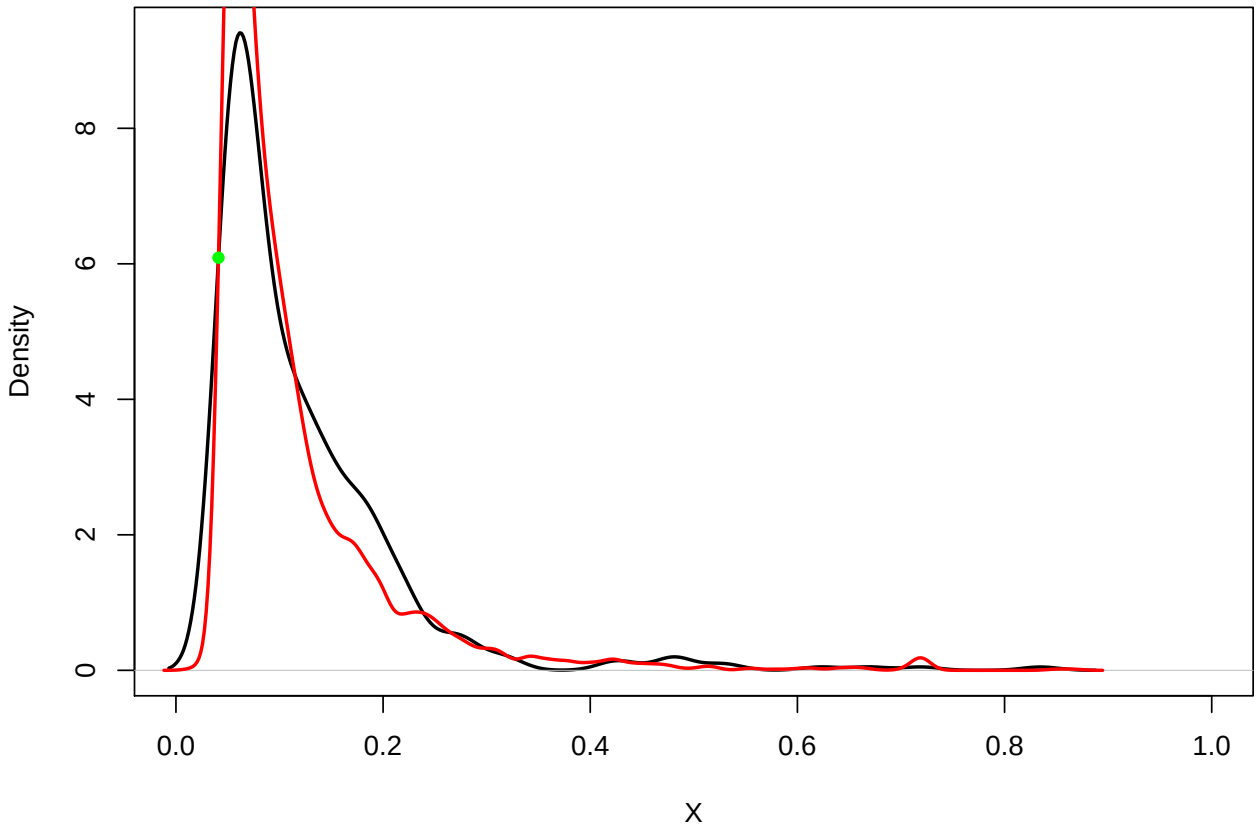

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2 Cluster 3
##       9856      1507      3327
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2 Cluster 3
##       72.4      77.8      70.5
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2 Cluster 3
##   232918.4  222207.6  243921.7
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2 Cluster 3
##        4302       688      1448
## B      4086       609      1420
## D      1110       159       349
## K       358        51       110
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2 Cluster 3
##      0.228     0.233     0.226
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2 Cluster 3
## Alberta                     320        45        93
## BritishColumbia             487        79       157
## NewBrunswick                 86        10        26
## NorthwestTerritories          5         0         2
## NovaScotia                  293        35       122
## Ontario                    1459       228       553
## Quebec                      596        76       186
## Saskatchewan                 47         6        18
## NA's                       6563      1028      2170
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2 Cluster 3
## 0      8905      1358      3031
## 1       743       112       229
## 2       100        18        29
## F       108        19        38
```

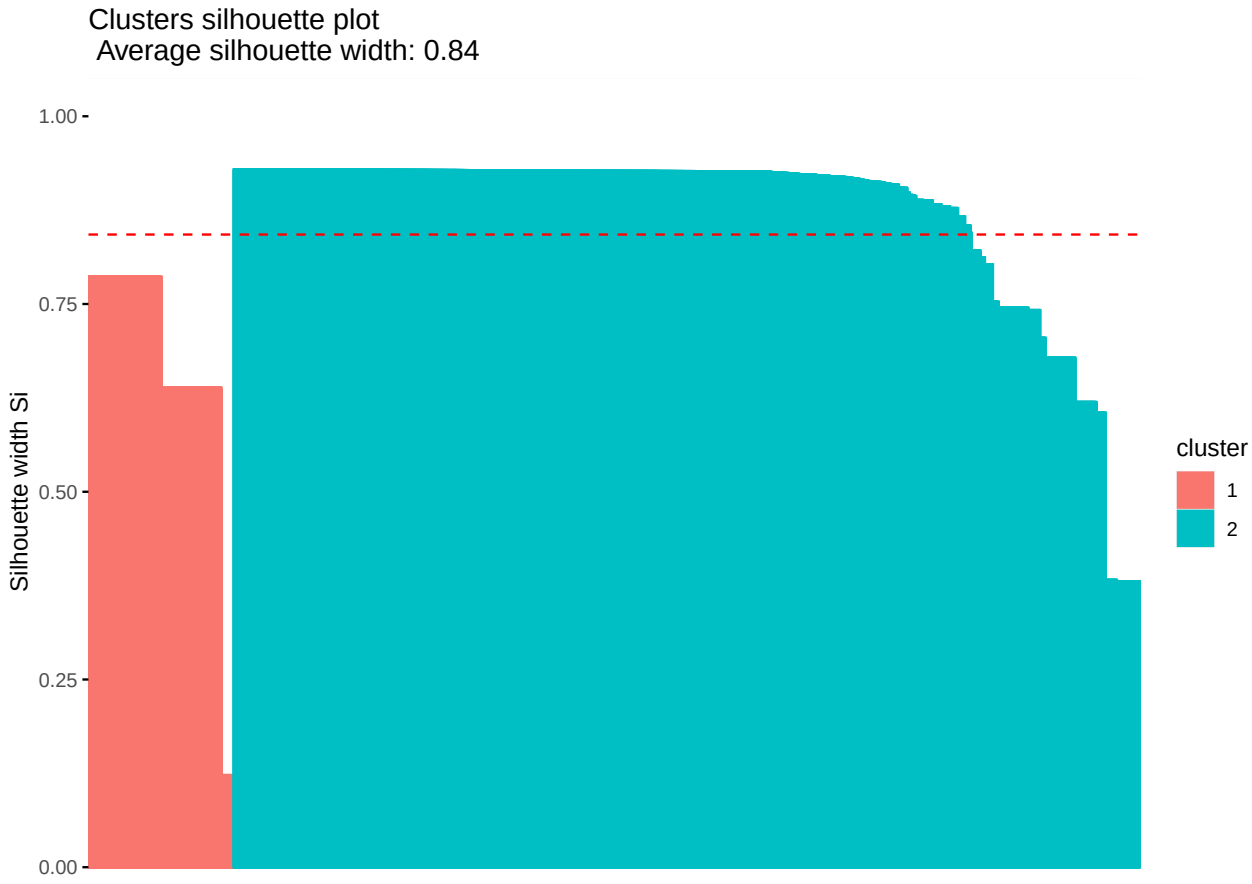
Libraries


Silhouette Coefficient
0.55
0.60
0.65
0.70
0.75
2
3
4
5
6
Number of Clusters


Cluster plot
Dim2 (20.3%)
Dim1 (35.9%)
5.0
2.5
0.0
−2.5
−10
−5
0
cluster
1
2
160209
145638
127906
307473
106850
418394
226317
293873
133775
199253
152739
218353
279167
104553
117395
18197
352842
4658
174166
177842
177656
168768
243150
184615
178382
177783
180461
185967
237490
471614



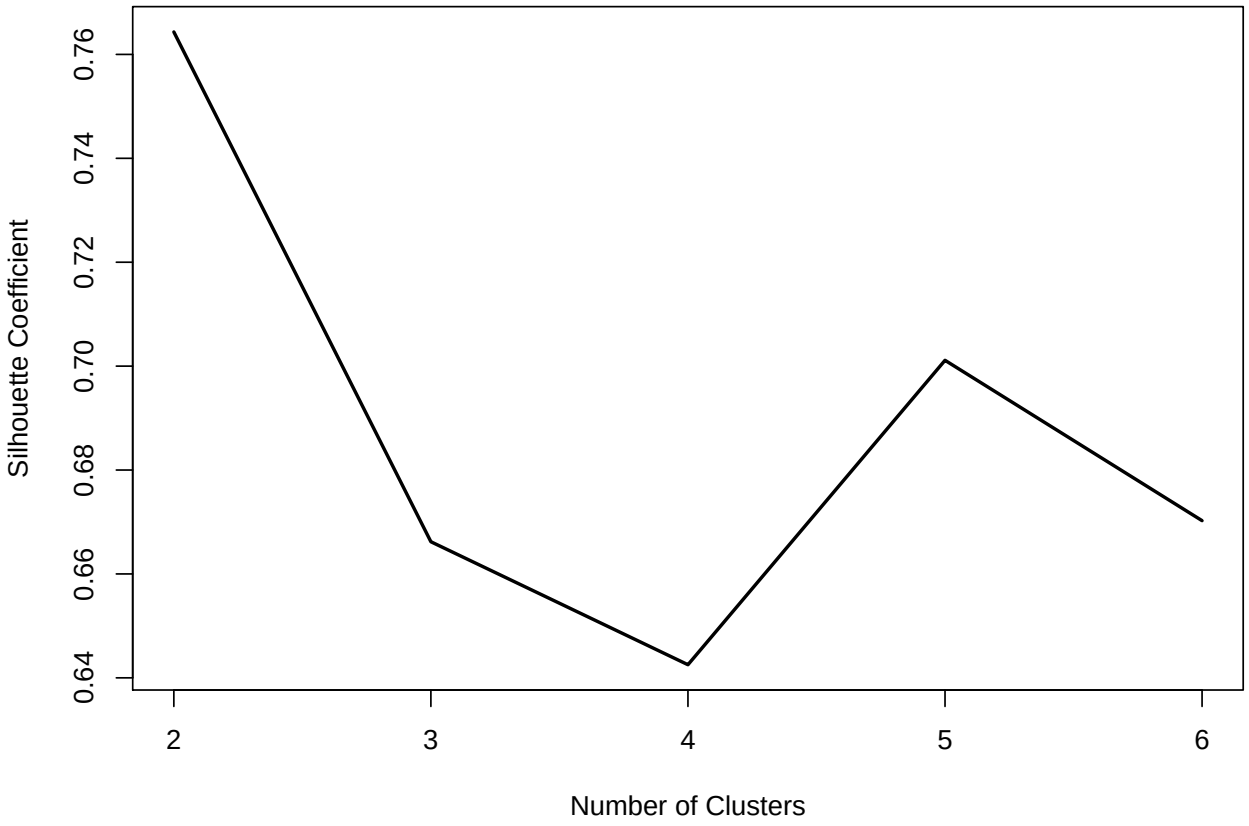

```
## [1] "Segment cutoff values:"
## [1] 0.04110552
##   cluster size ave.sil.width
## 1       1  454          0.67
## 2       2 2836          0.87
```

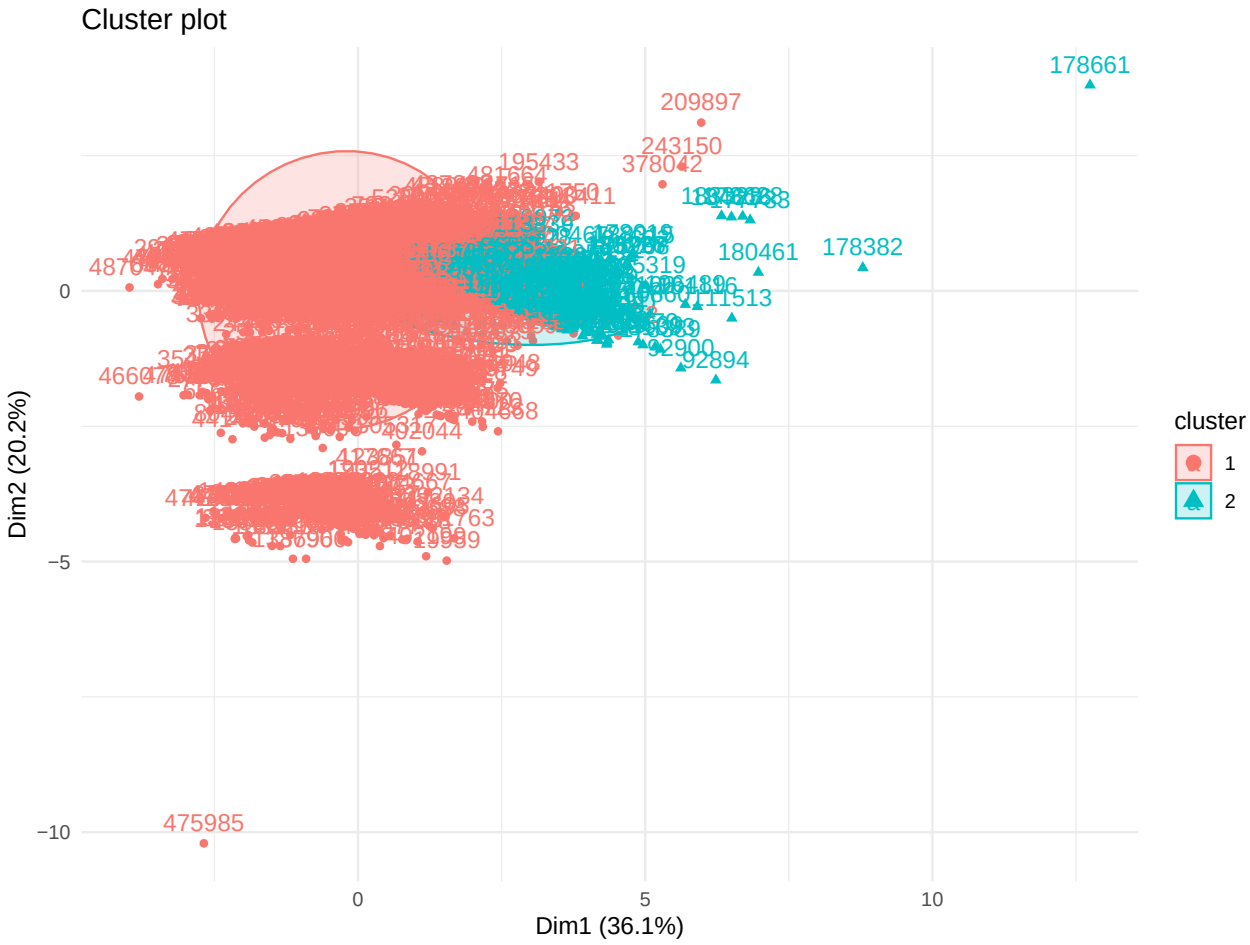


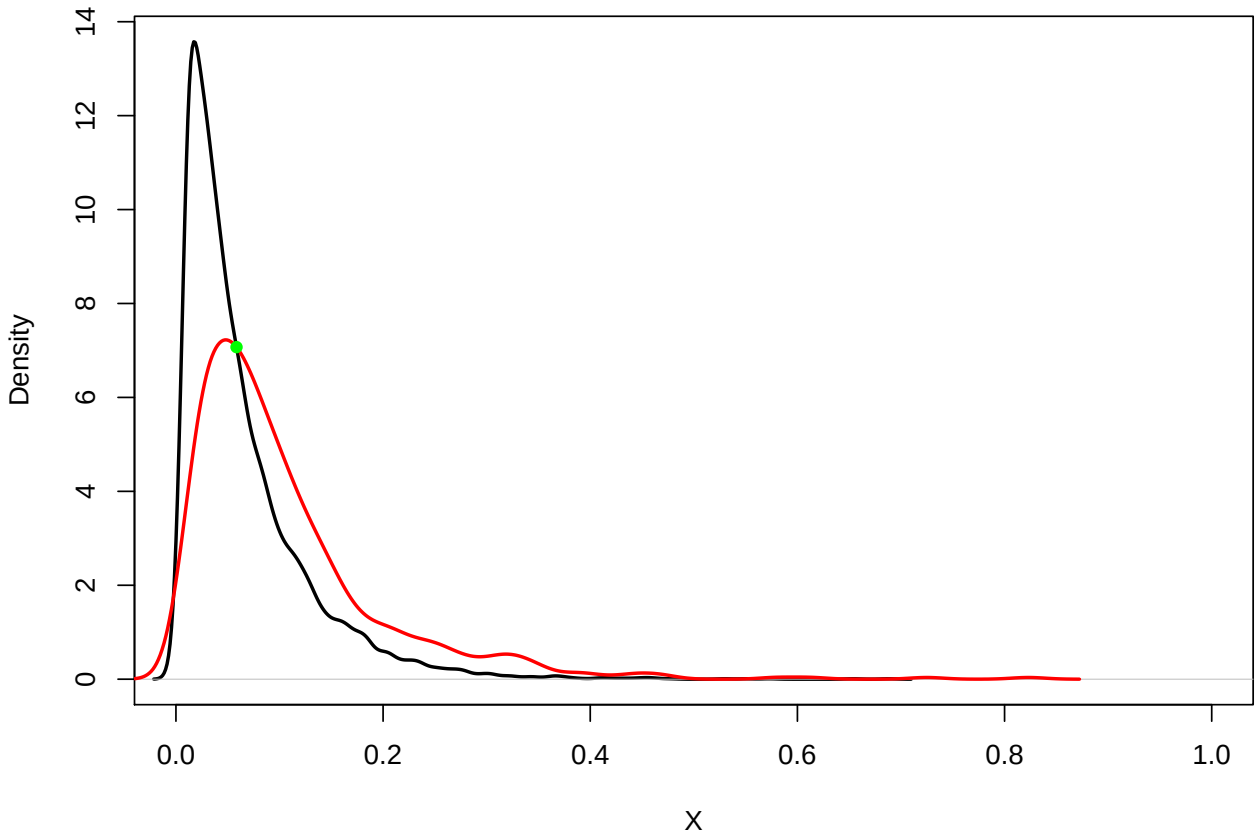

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##       2037     12653
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       71.2      72.7
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   232758.6  234560.4
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2
##         903      5535
## B       830      5285
## D       229      1389
## K        75       444
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2
##      0.233     0.228
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2
## Alberta                      56       402
## BritishColumbia              96       627
## NewBrunswick                 20       102
## NorthwestTerritories          1         6
## NovaScotia                   54       396
## Ontario                     297      1943
## Quebec                      124       734
## Saskatchewan                 11        60
## NA's                       1378      8383
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0      1840     11454
## 1       154       930
## 2        25       122
## F        18       147
```

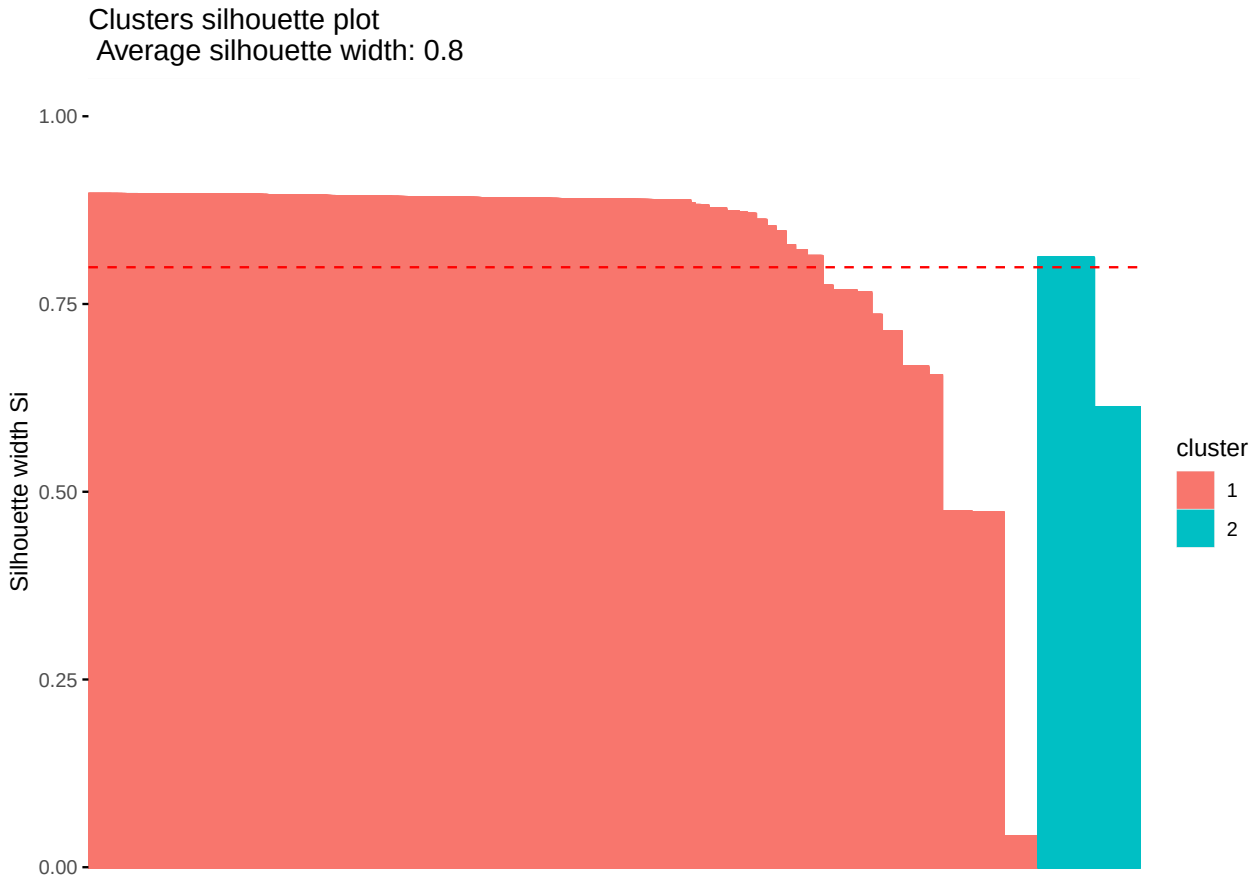
Parks


Silhouette Coefficient
0.64
0.66
0.68
0.70
0.72
0.74
0.76
2
3
4
5
6
Number of Clusters


Cluster plot
178661
209897
243150
378042
195433
180461
178382
475985
Dim2 (20.2%)
Dim1 (36.1%)
0
−5
−10
0
5
10
cluster
1
2



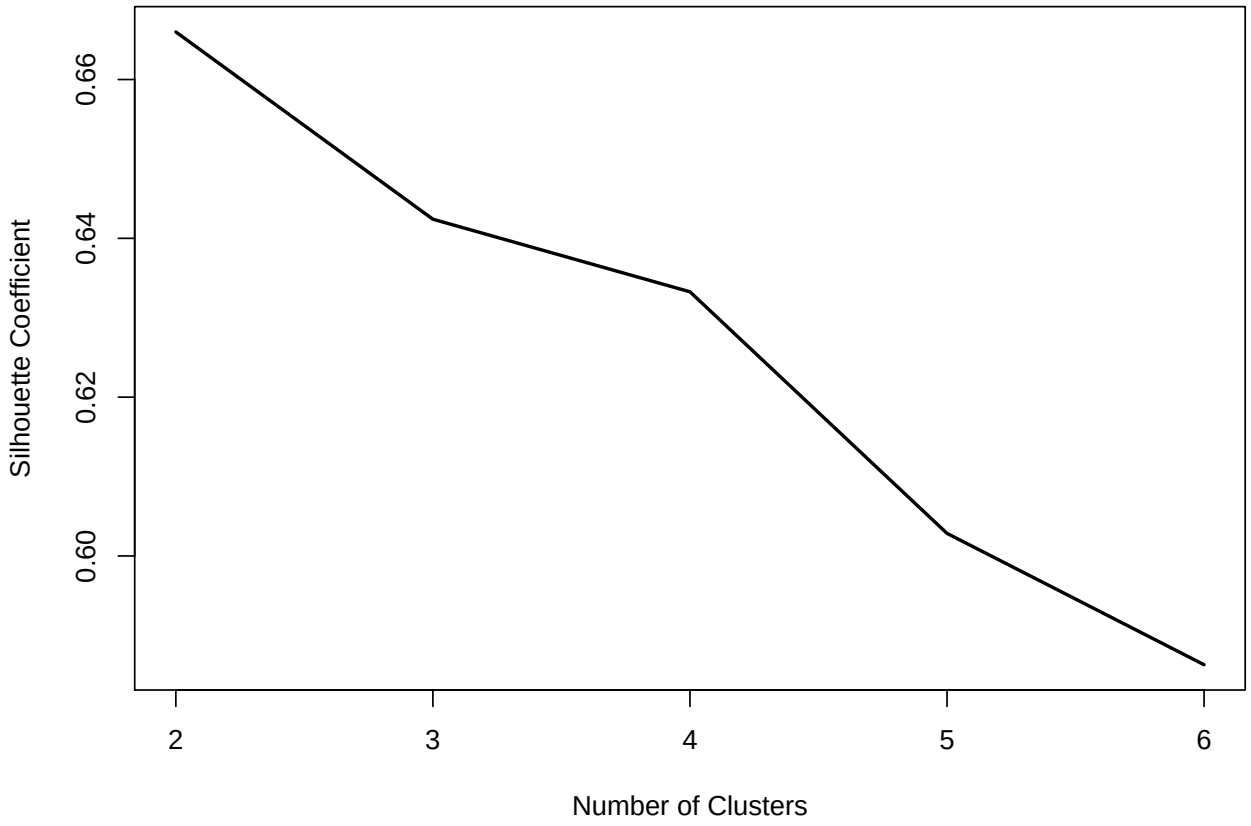

```
## [1] "Segment cutoff values:"
## [1] 0.0585062
##   cluster size ave.sil.width
## 1       1 6296          0.81
## 2       2  673          0.72
```

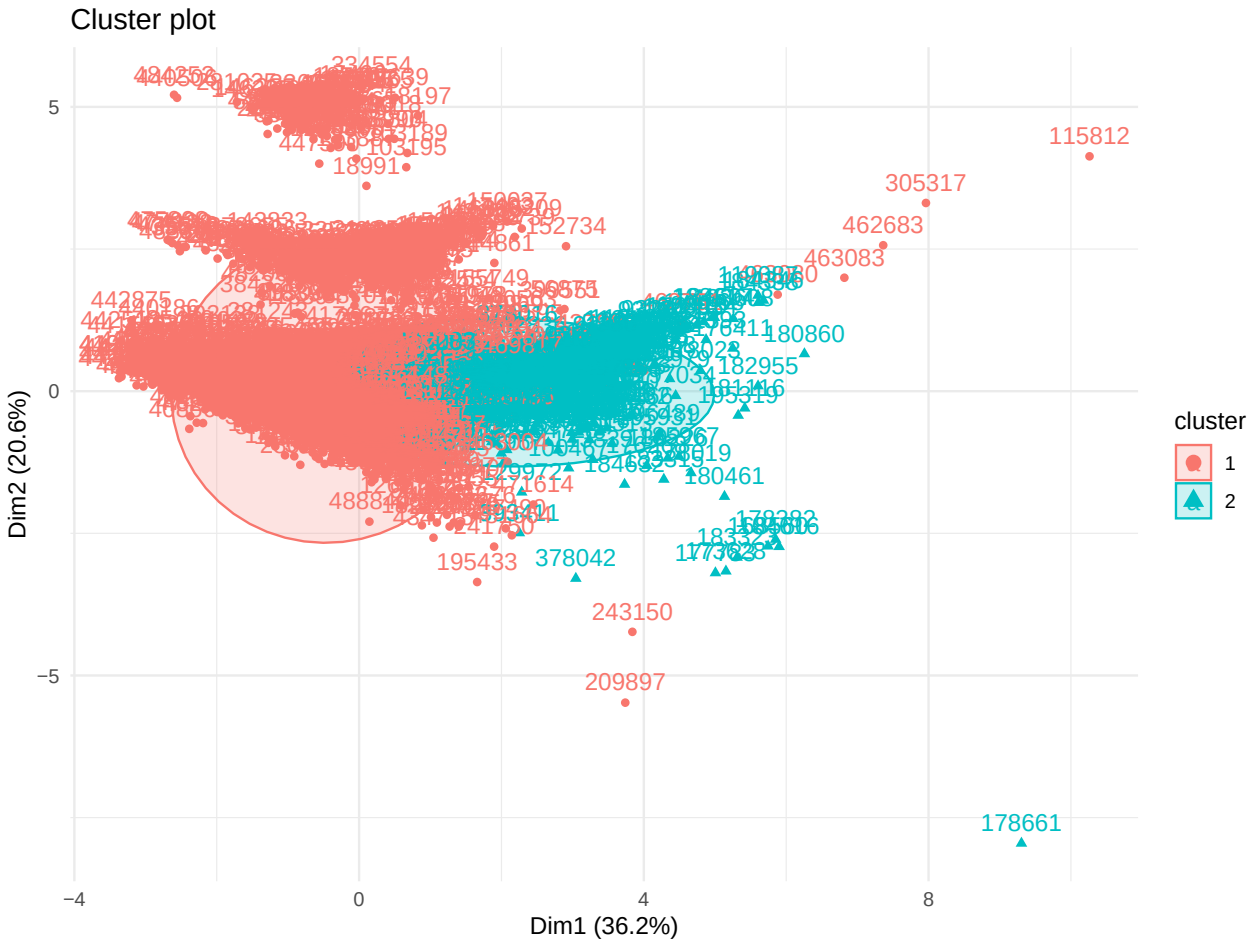


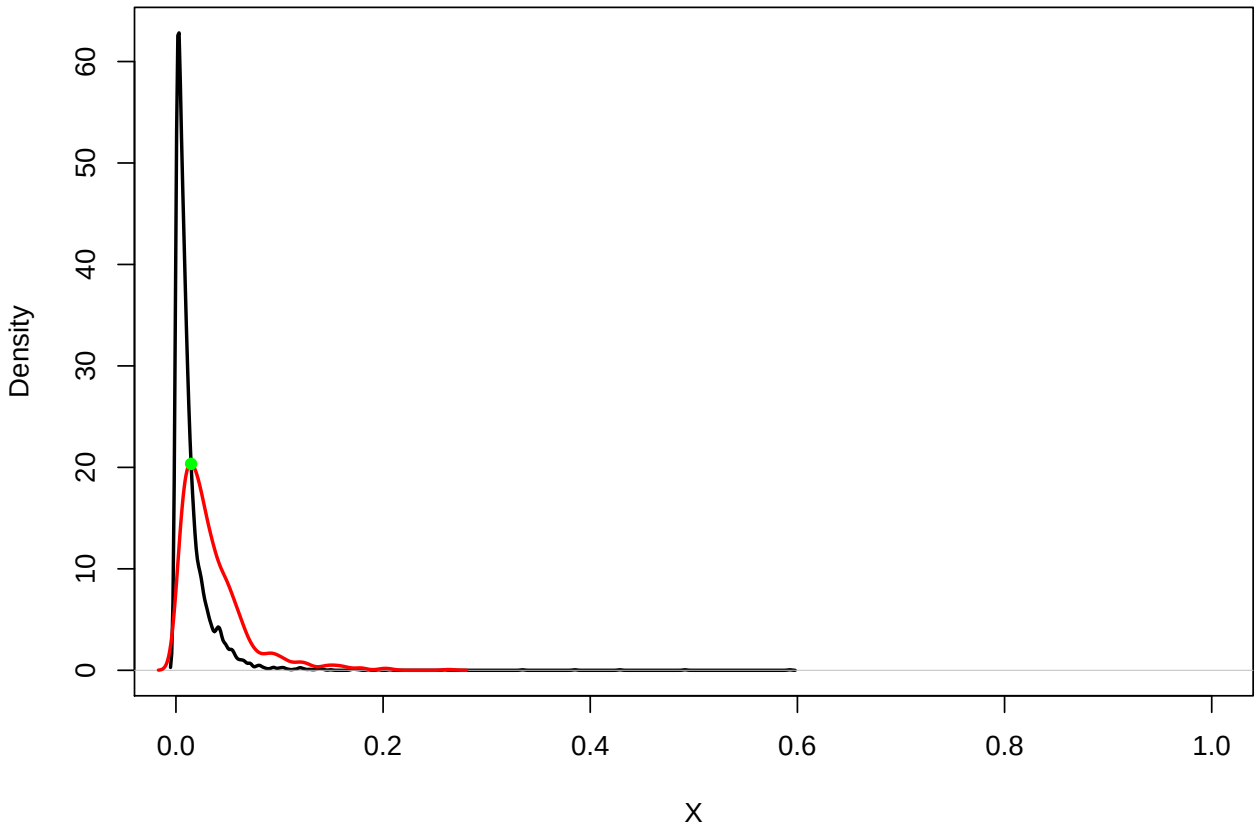

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##      13271      1419
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       73.1      66.9
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   237127.2    207982
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2
##        5802       636
## B      5551       564
## D      1445       173
## K       473        46
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2
##      0.228     0.233
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2
## Alberta                     416        42
## BritishColumbia             654        69
## NewBrunswick                112        10
## NorthwestTerritories          7         0
## NovaScotia                  399        51
## Ontario                    2028       212
## Quebec                      785        73
## Saskatchewan                 63         8
## NA's                       8807       954
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0     11994      1300
## 1       999        85
## 2       132        15
## F       146        19
```

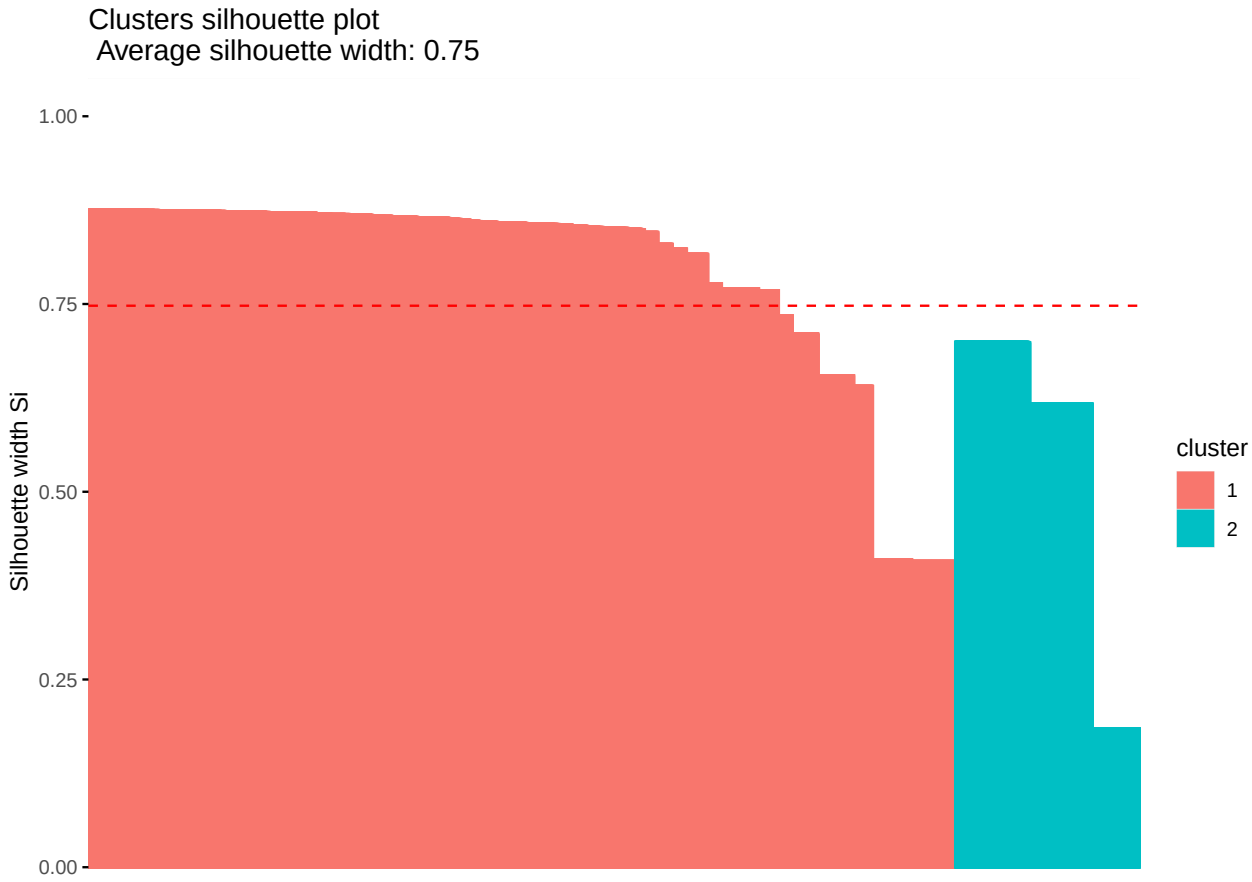
Transit


Silhouette Coefficient
0.60
0.62
0.64
0.66
2
3
4
5
6
Number of Clusters


Cluster plot
Dim2 (20.6%)
Dim1 (36.2%)
5
0
−5
−4
0
4
8
cluster
1
2
115812
305317
462683
463083
180860
182955
180461
378042
243150
209897
178661
195433




```
## [1] "Segment cutoff values:"
## [1] 0.01464926
##   cluster size ave.sil.width
## 1       1 4411          0.79
## 2       2  941          0.54
```




```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##      12088      2602
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       72.6      72.2
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##     231179  248852.3
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2
##        5287      1151
## B      5014      1101
## D      1338       280
## K       449        70
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2
##      0.228     0.228
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2
## Alberta                     385        73
## BritishColumbia             592       131
## NewBrunswick                102        20
## NorthwestTerritories          7         0
## NovaScotia                  362        88
## Ontario                    1825       415
## Quebec                      710       148
## Saskatchewan                 59        12
## NA's                       8046      1715
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0     10919      2375
## 1       907       177
## 2       123        24
## F       139        26
```

Conclusion

text