k-means with Imputation

`ClustImpute` package

PMS

17 May, 2023

Assumptions of the Alogrithm

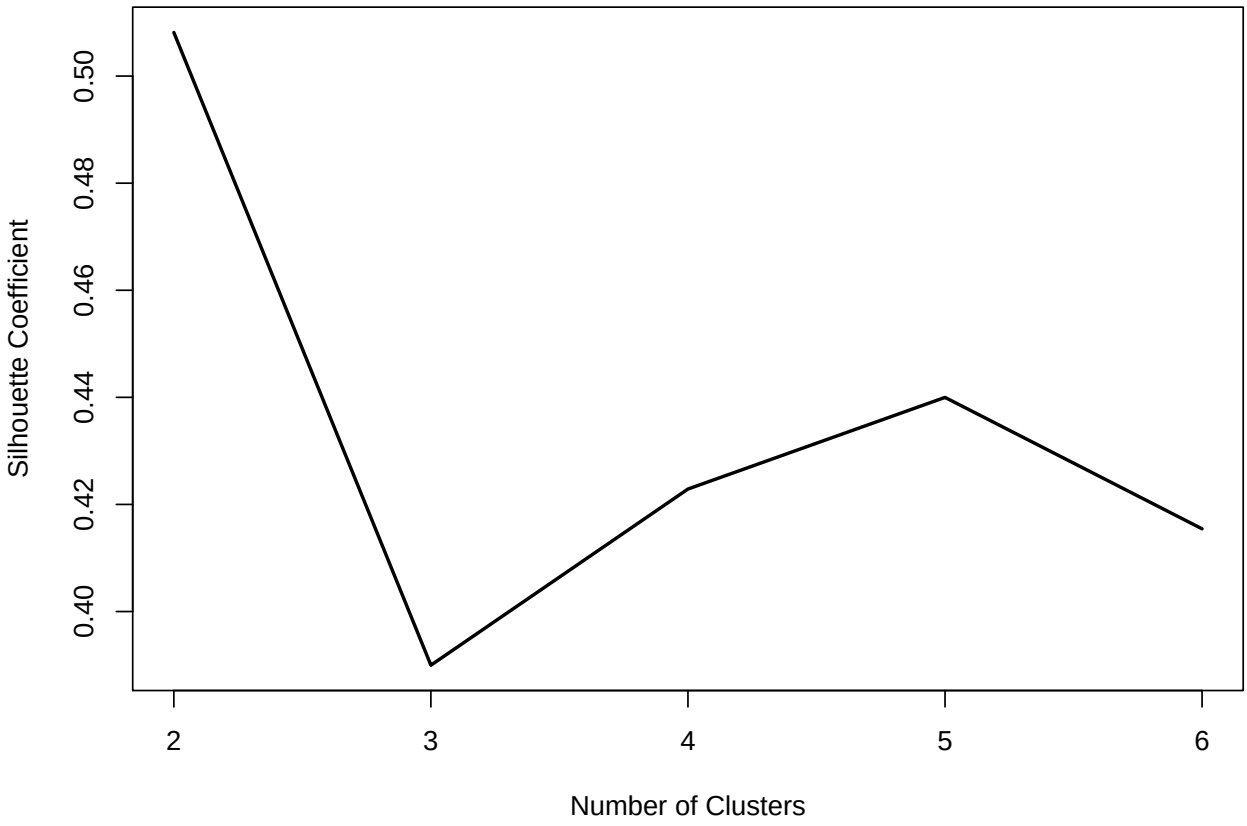
This algorithm "draws the missing values iteratively based on the current cluster assignment so that correlations are considered on this level". Also, "penalizing weights are imposed on imputed values and successively decreased (to zero) as the missing data imputation gets better". The idea is that the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

Algorithm steps:

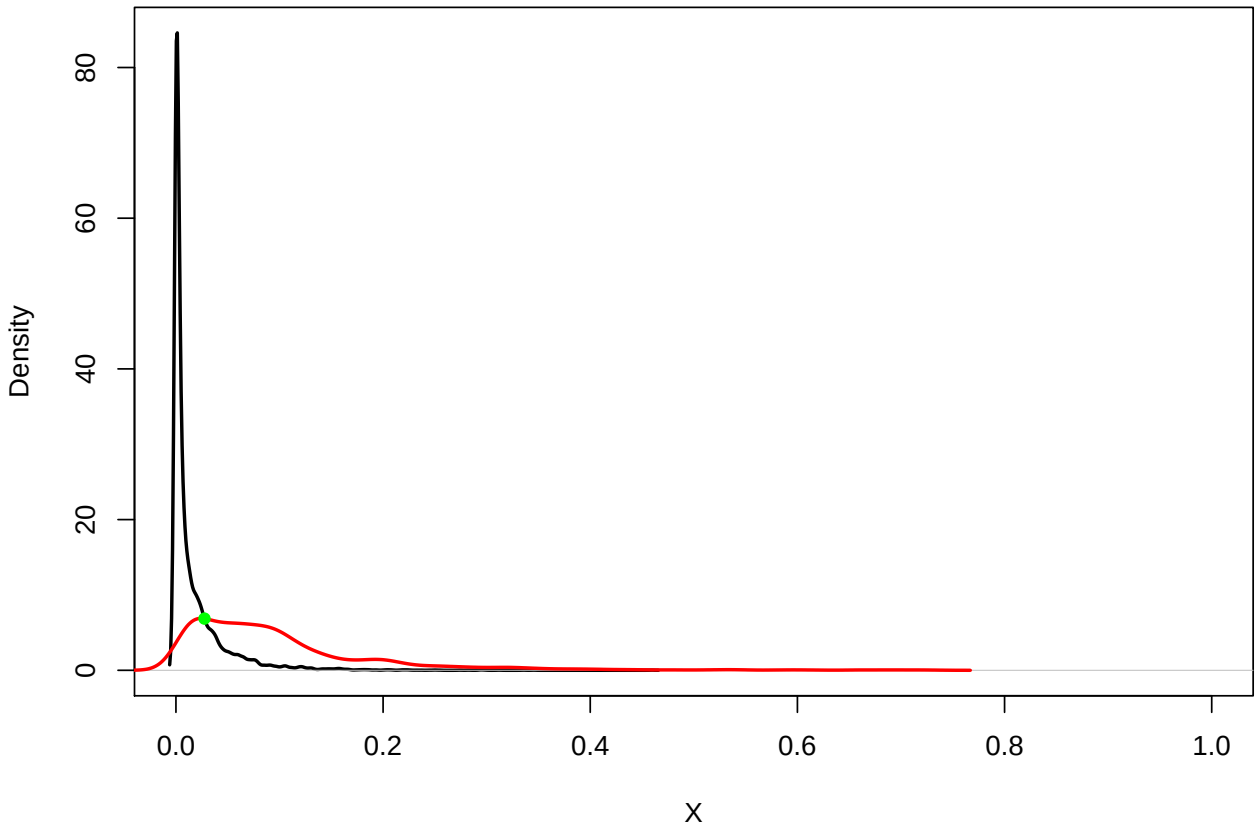
1. It replaces all NAs by random imputation, i.e., for each variable with missings, it draws from the marginal distribution of this variable not taking into account any correlations with other variables
2. Weights < 1 are used to adjust the scale of an observation that was generated in step 1. The weights 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

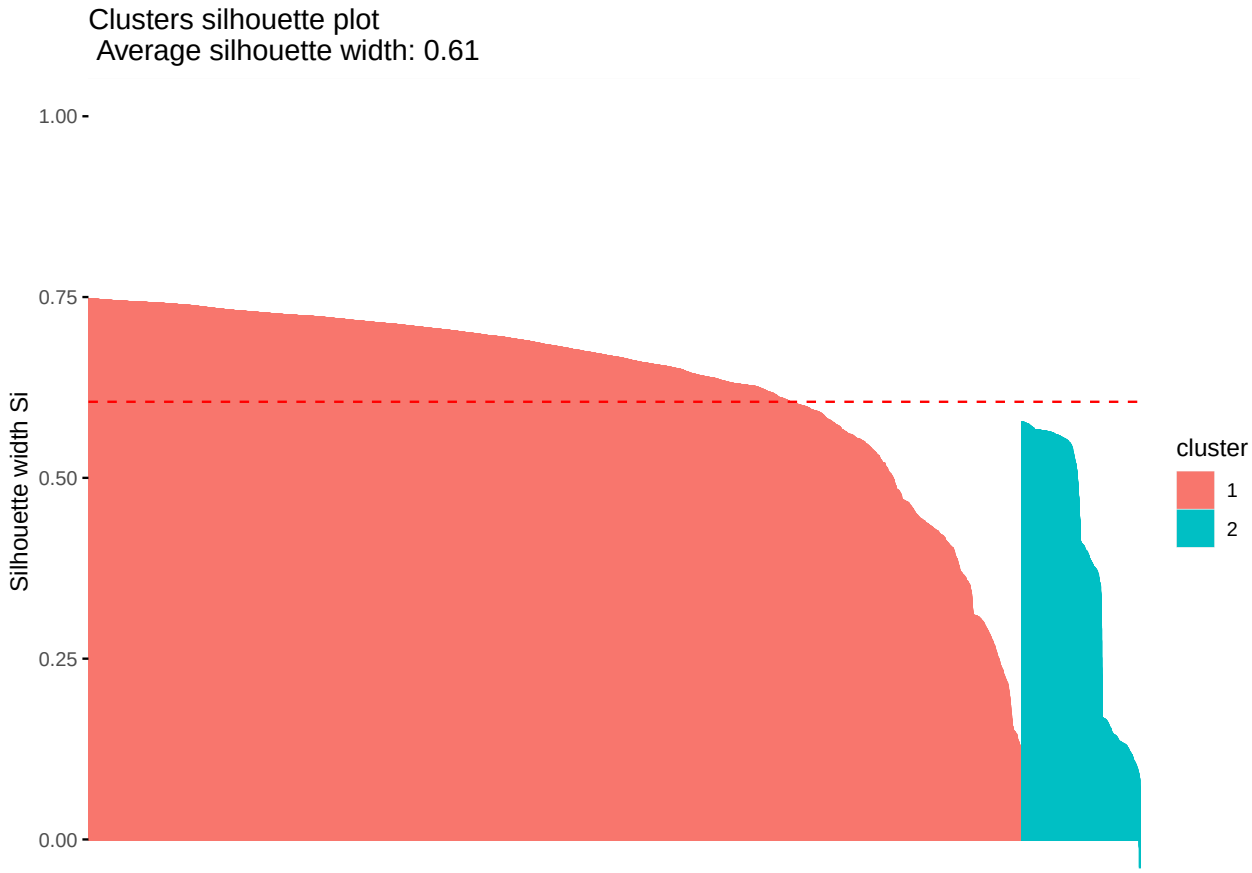
Employment




Cluster plot
Dim2 (20.1%)
Dim1 (43.5%)
cluster
1
2
8
4
0
0
5
10
15
310724
178661
243150
209897
378042



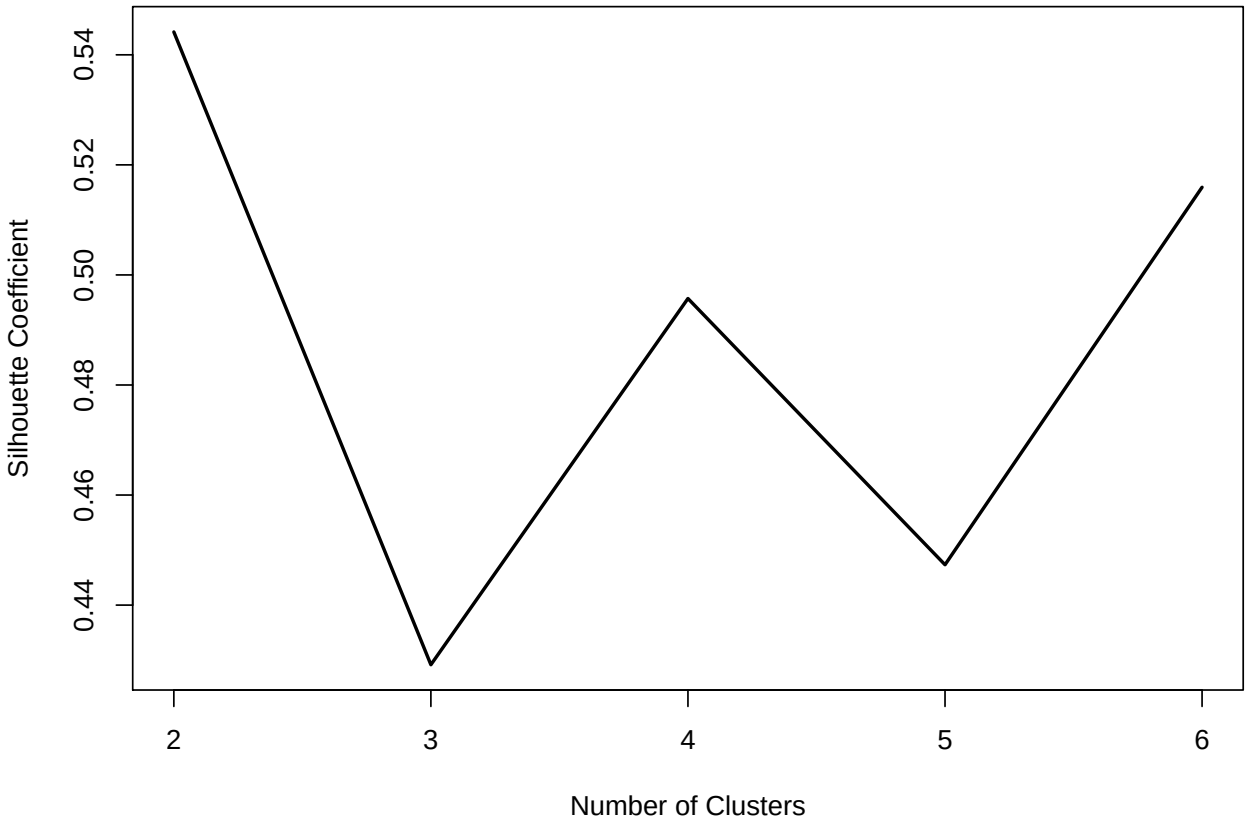

```
## [1] "Segment cutoff values:"
## [1] 0.02756497
##   cluster  size ave.sil.width
## 1       1 11236          0.63
## 2       2  1417          0.38
```

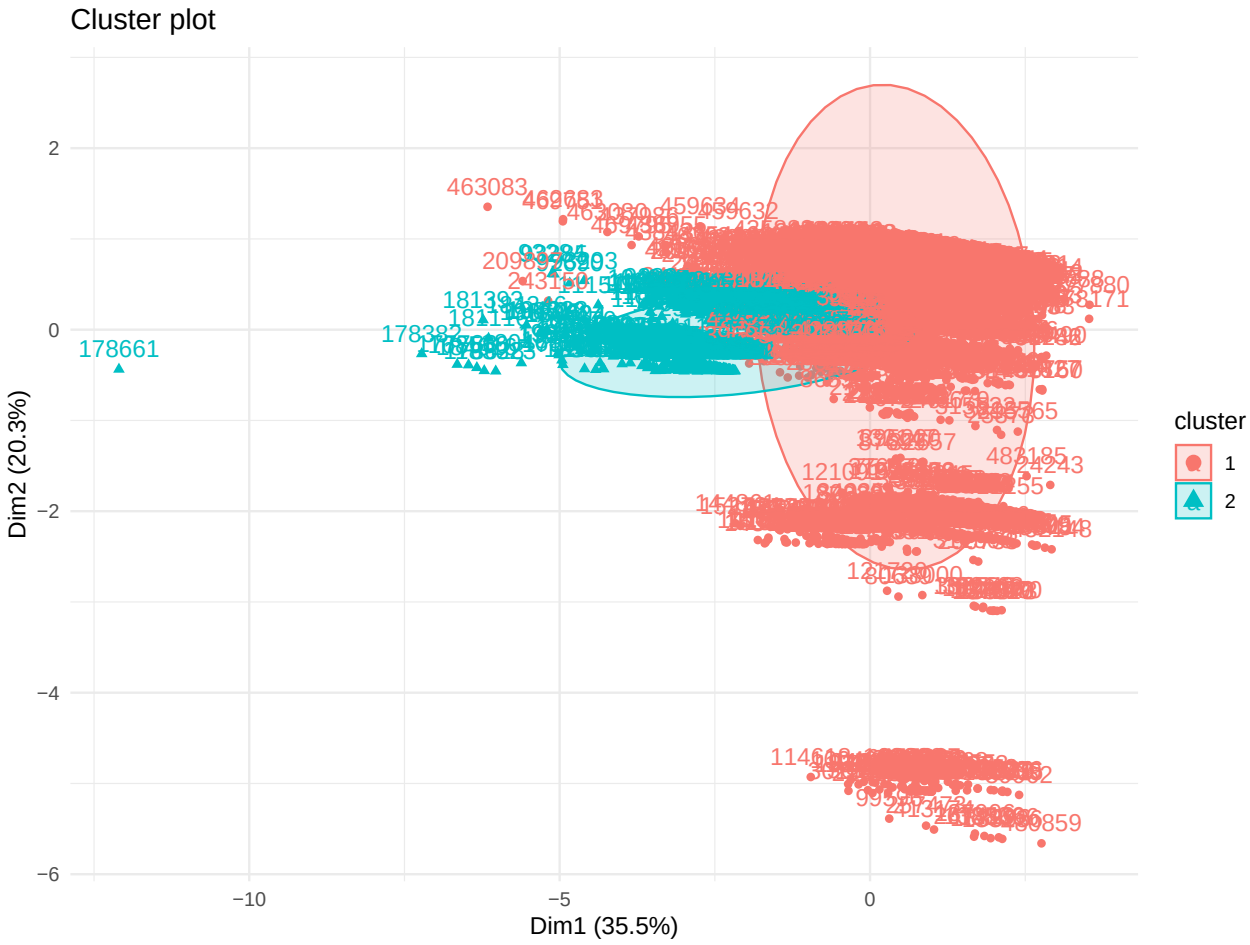


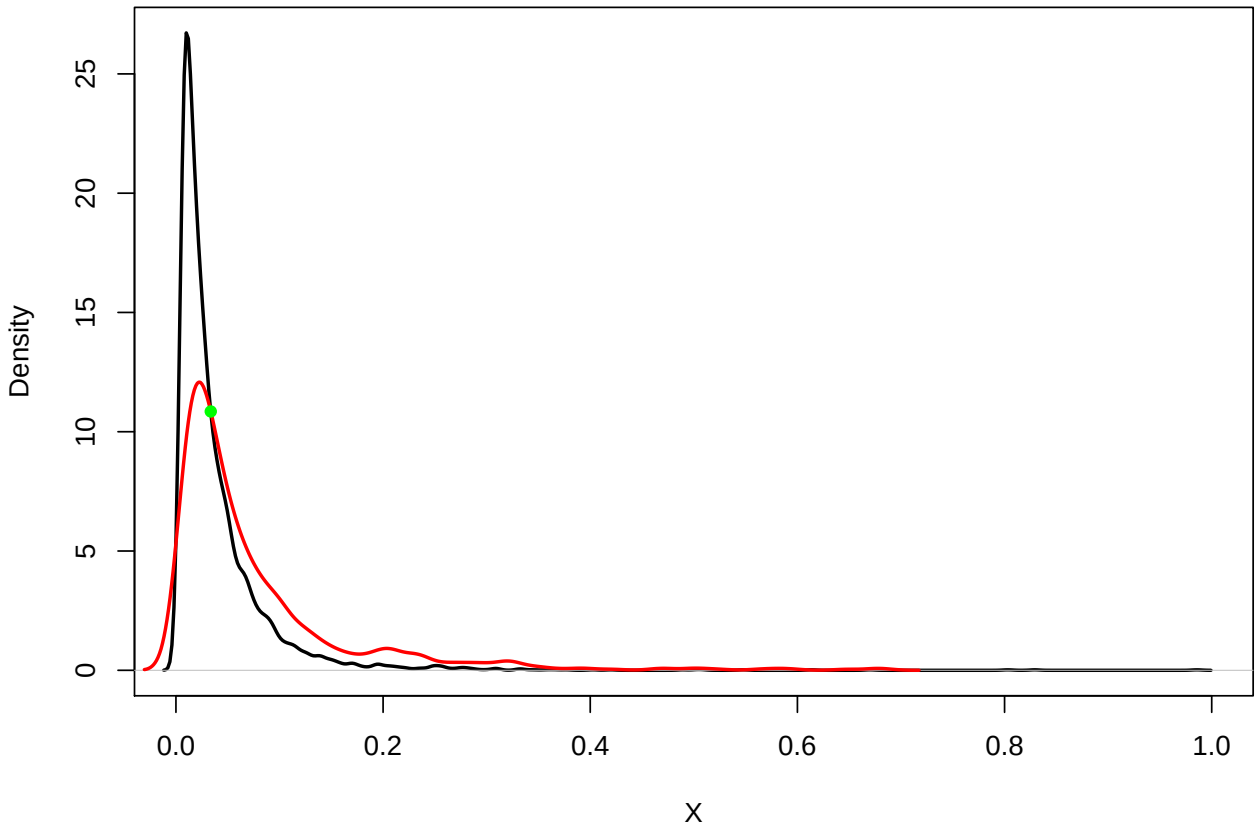

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##      13055      1635
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       72.6      72.3
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   235151.2  227602.3
##
##
##
##  CMA Type:
##    Cluster 1 Cluster 2
##         5736       702
## B       5427       688
## D       1433       185
## K        459        60
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2
##      0.228      0.23
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2
## Alberta                     408        50
## BritishColumbia             652        71
## NewBrunswick                105        17
## NorthwestTerritories          7         0
## NovaScotia                  402        48
## Ontario                    1986       254
## Quebec                      772        86
## Saskatchewan                 67         4
## NA's                       8656      1105
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0     11810      1484
## 1       960       124
## 2       137        10
## F       148        17
```

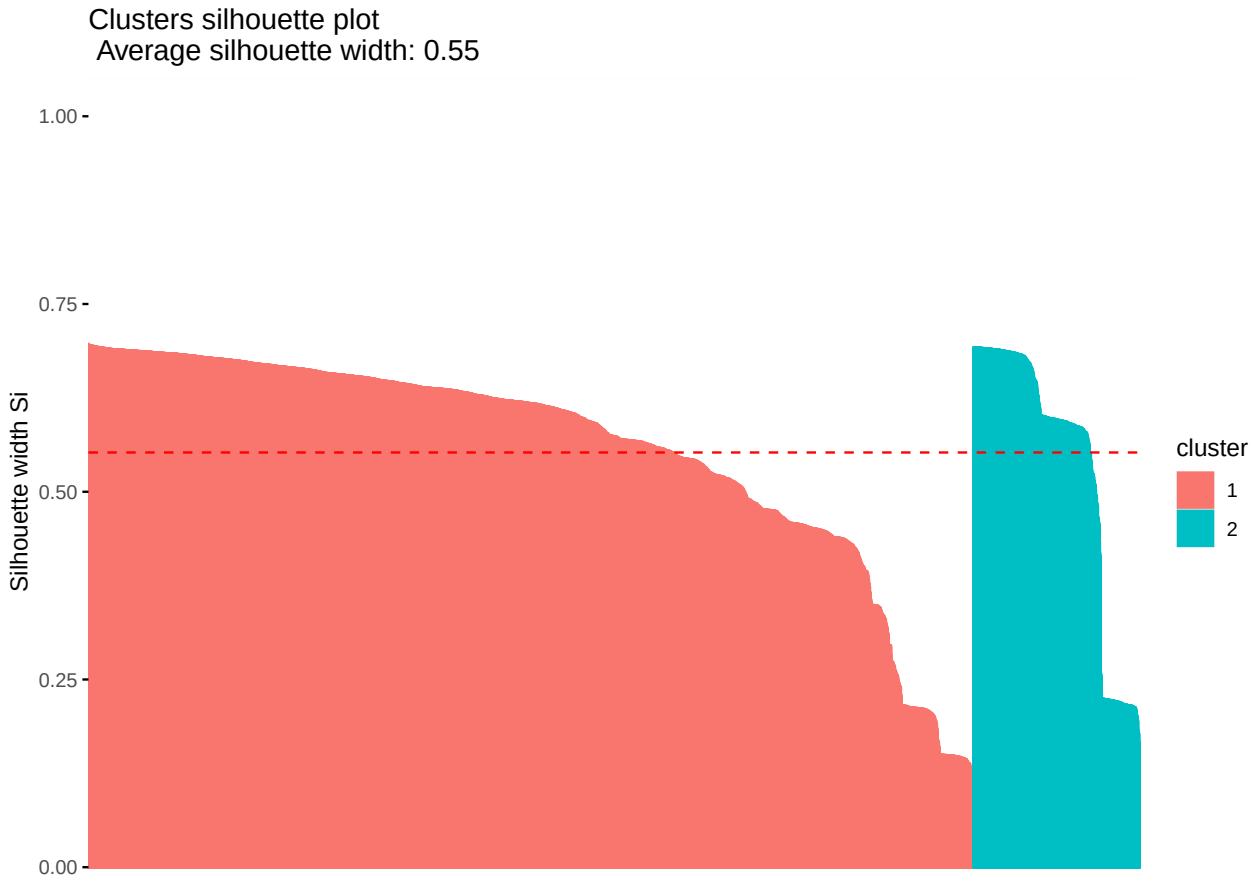
Pharmacy


Silhouette Coefficient
0.44
0.46
0.48
0.50
0.52
0.54
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



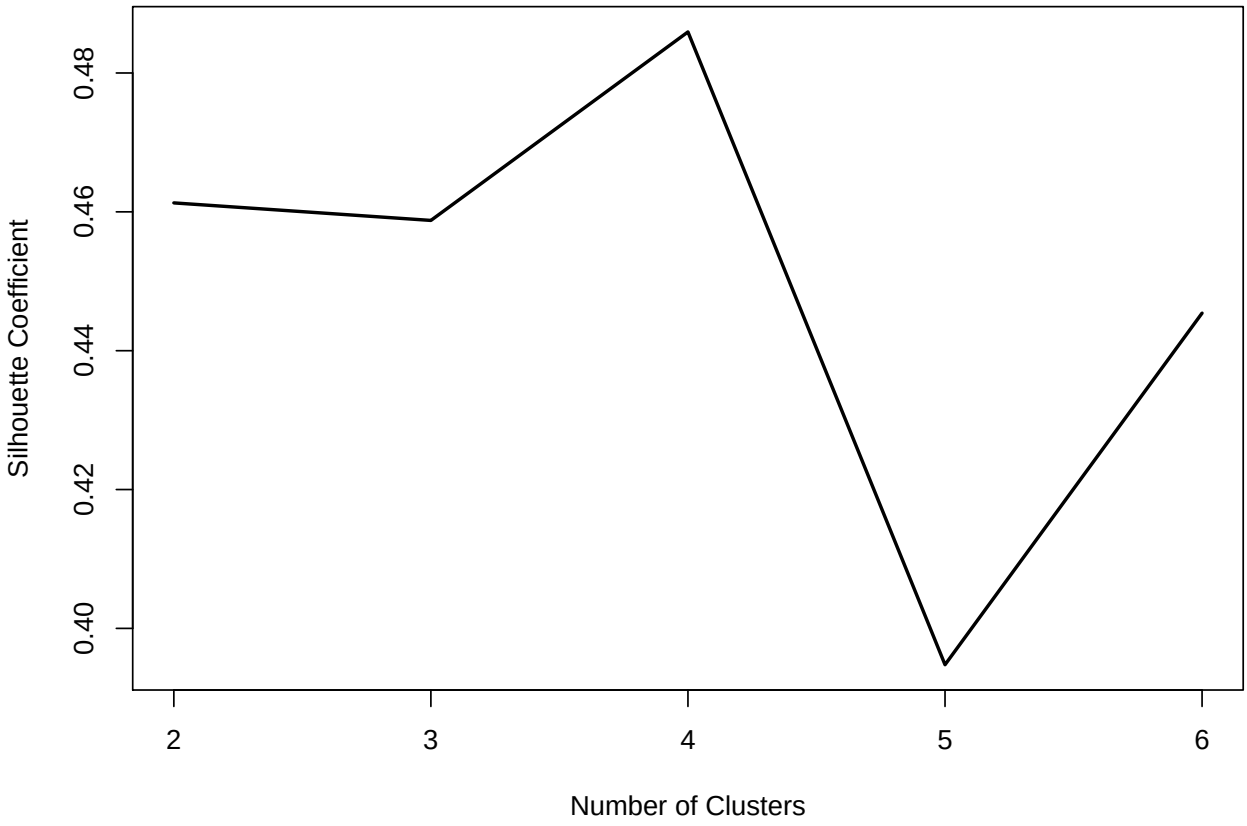

```
## [1] "Segment cutoff values:"
## [1] 0.03363201
##   cluster size ave.sil.width
## 1       1 4403          0.56
## 2       2  828          0.53
```

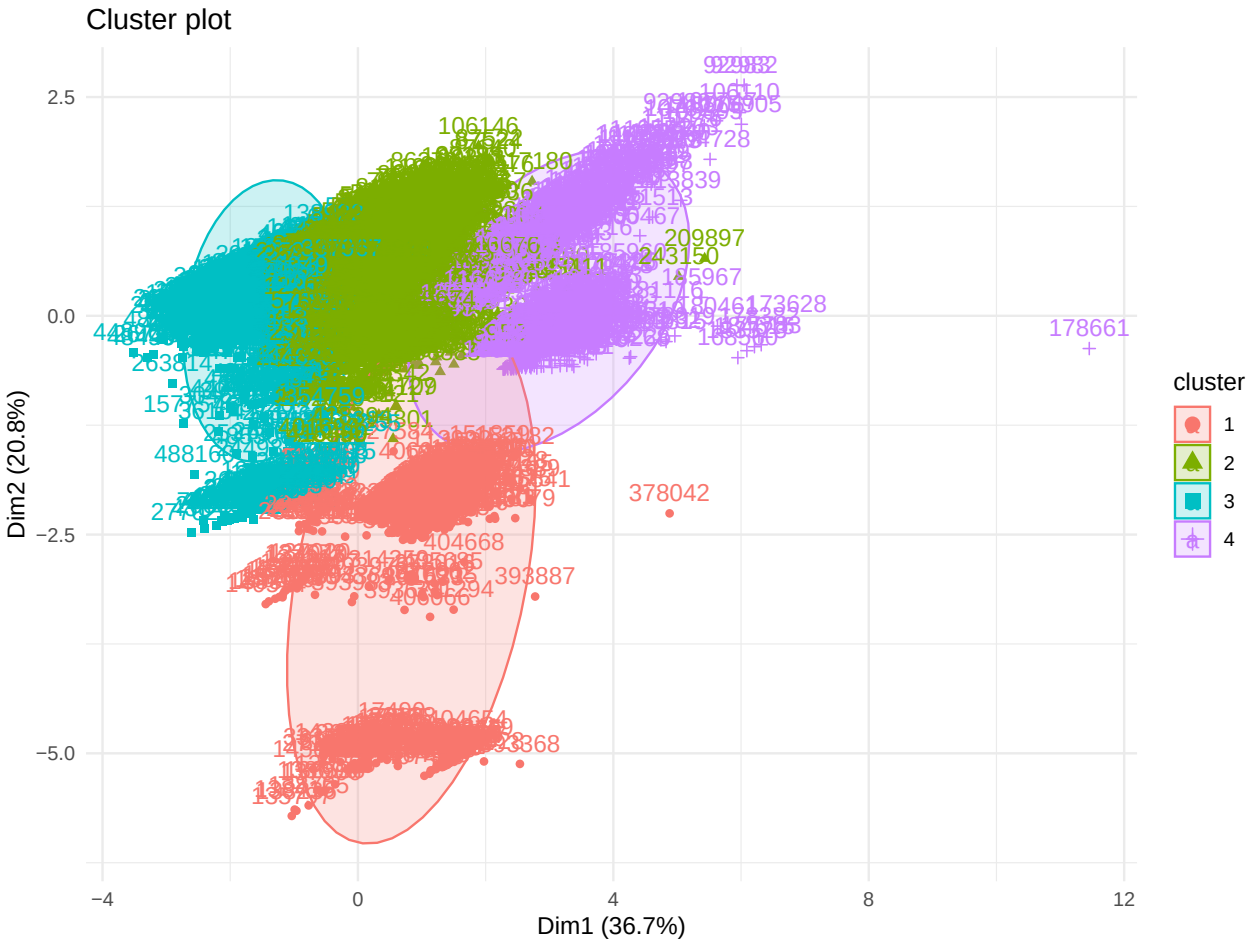


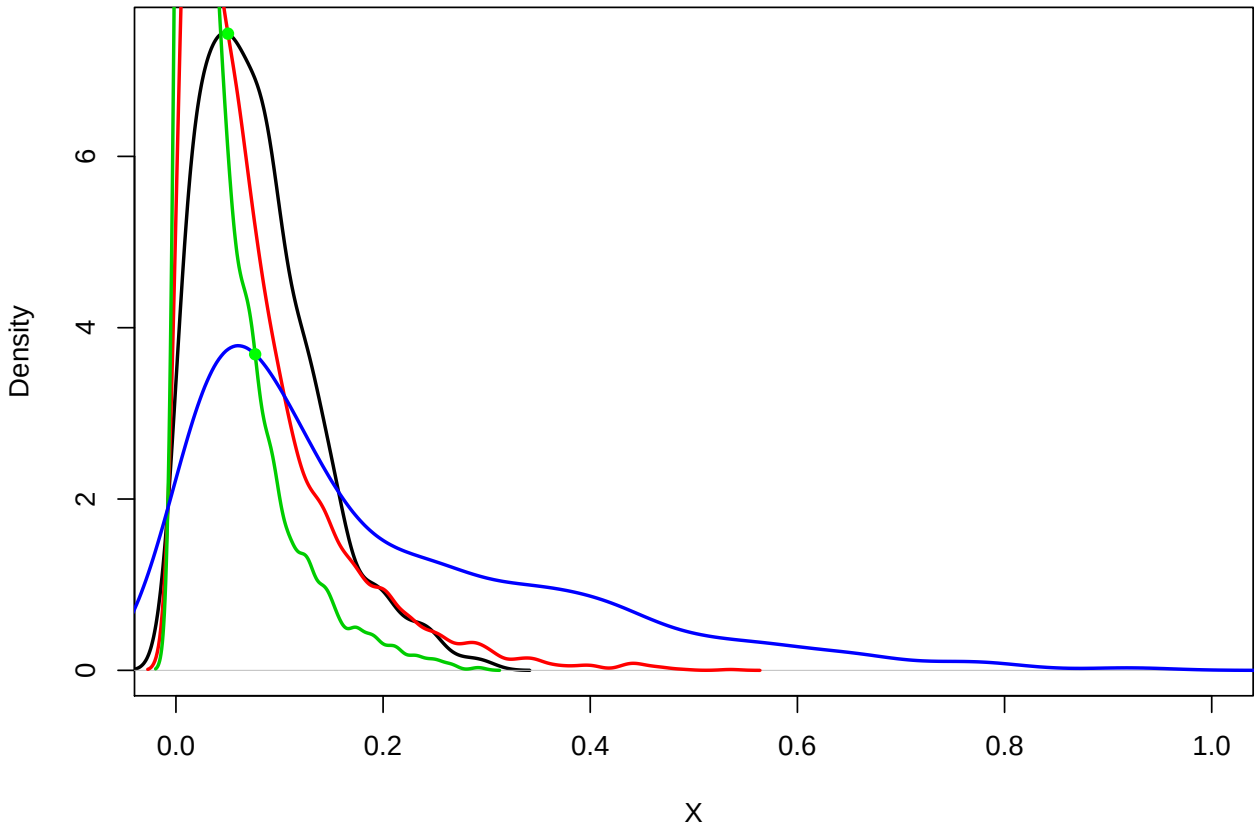

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##      12341      2349
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       73.7      66.3
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   233788.4  237056.2
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2
##        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
```

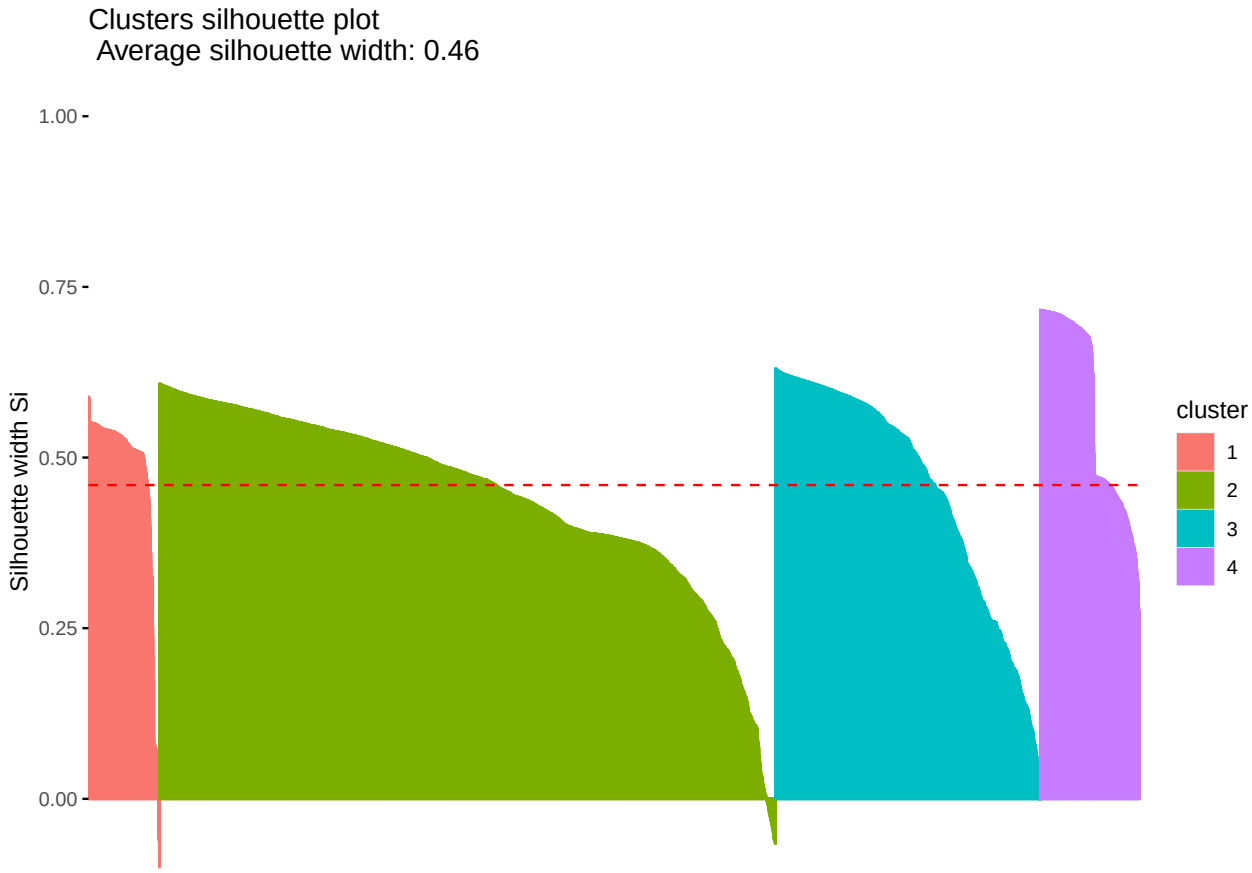
Childcare


Silhouette Coefficient
0.40
0.42
0.44
0.46
0.48
2
3
4
5
6
Number of Clusters


Cluster plot
Dim2 (20.8%)
Dim1 (36.7%)
2.5
0.0
−2.5
−5.0
−4
0
4
8
12
cluster
1
2
3
4
178661
378042
404668
393887
209897
243150
173628



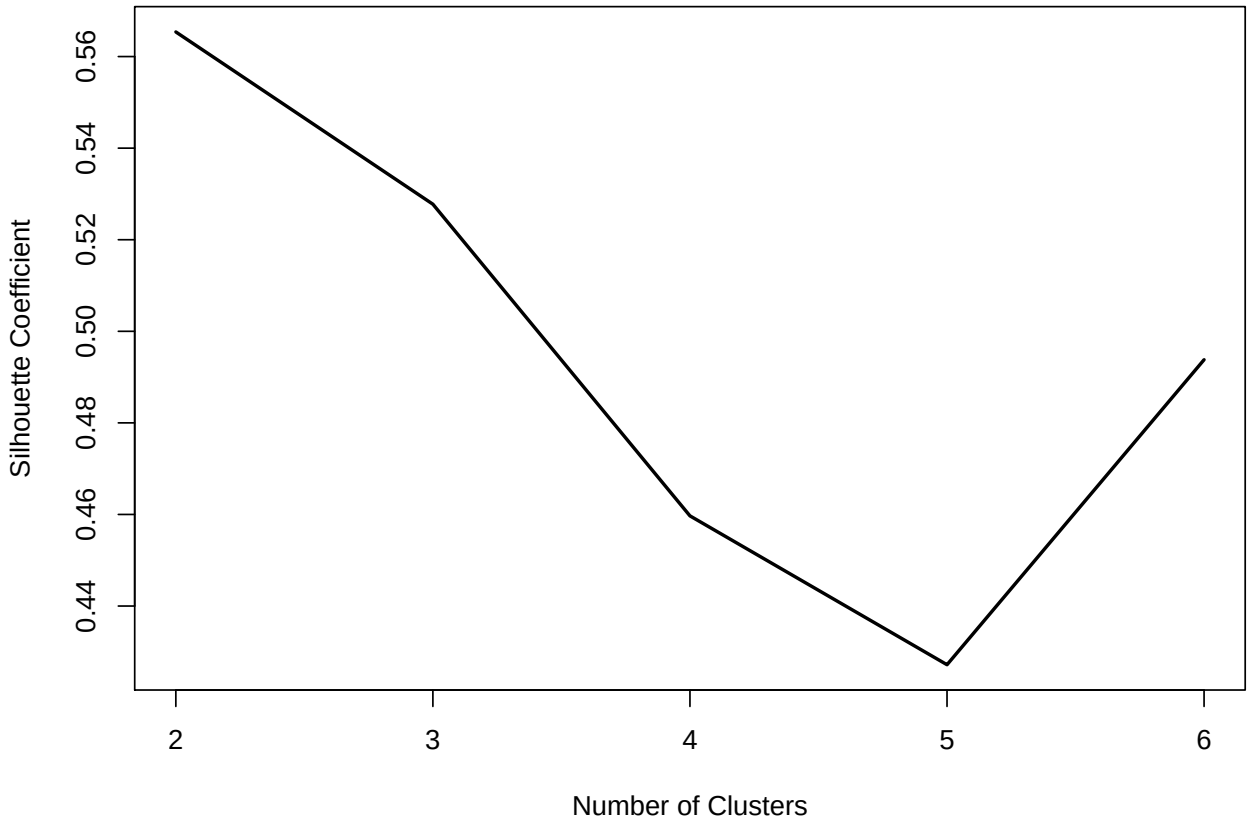

```
## [1] "Segment cutoff values:"
## [1] 0.05025143
## [1] 0.03861828
## [1] 0.07648809
##   cluster size ave.sil.width
## 1       1  489          0.47
## 2       2 4236          0.44
## 3       3 1819          0.45
## 4       4  687          0.57
```

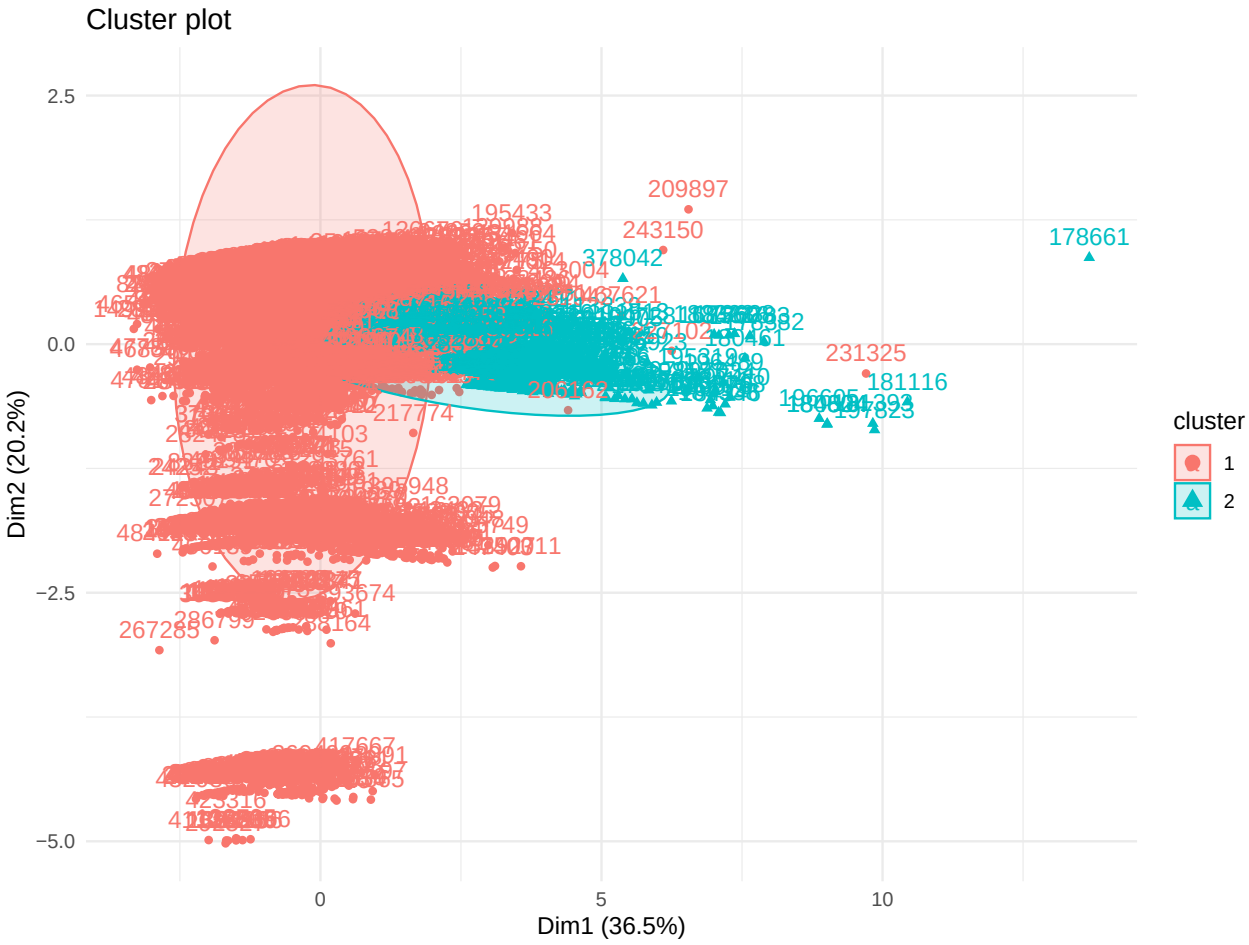


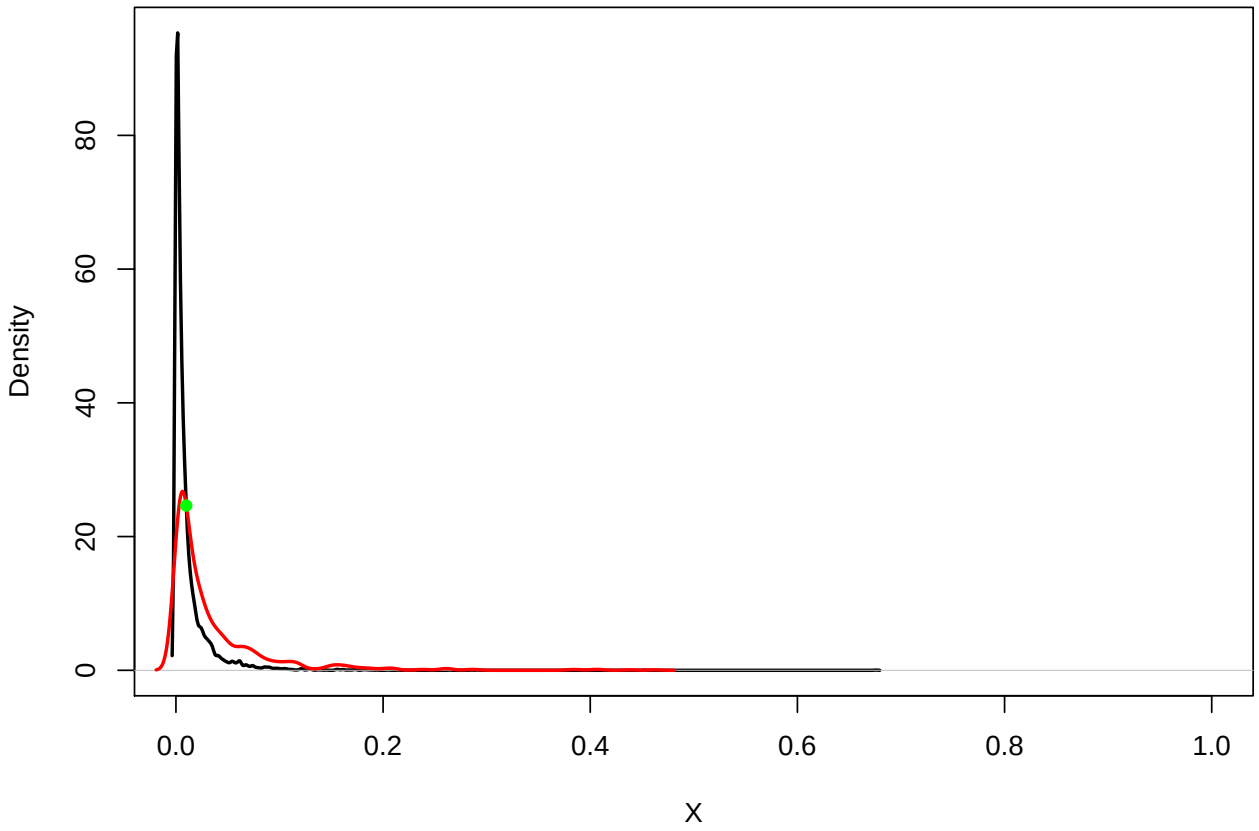

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##        995      8603      3696      1396
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##       70.9      73.7      71.1        70
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##     236803  230010.6    240763  241953.1
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2 Cluster 3 Cluster 4
##         462      3778      1587       611
## B       407      3553      1571       584
## D        99       963       407       149
## K        27       309       131        52
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##      0.237     0.229     0.224      0.23
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2 Cluster 3 Cluster 4
## Alberta                      34       263       119        42
## BritishColumbia              42       415       204        62
## NewBrunswick                  2        74        34        12
## NorthwestTerritories          0         4         3         0
## NovaScotia                   31       273       113        33
## Ontario                     139      1319       564       218
## Quebec                       63       525       207        63
## Saskatchewan                  5        41        18         7
## NA's                        679      5689      2434       959
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2 Cluster 3 Cluster 4
## 0       893      7816      3333      1252
## 1        79       621       271       113
## 2        12        82        42        11
## F        11        84        50        20
```

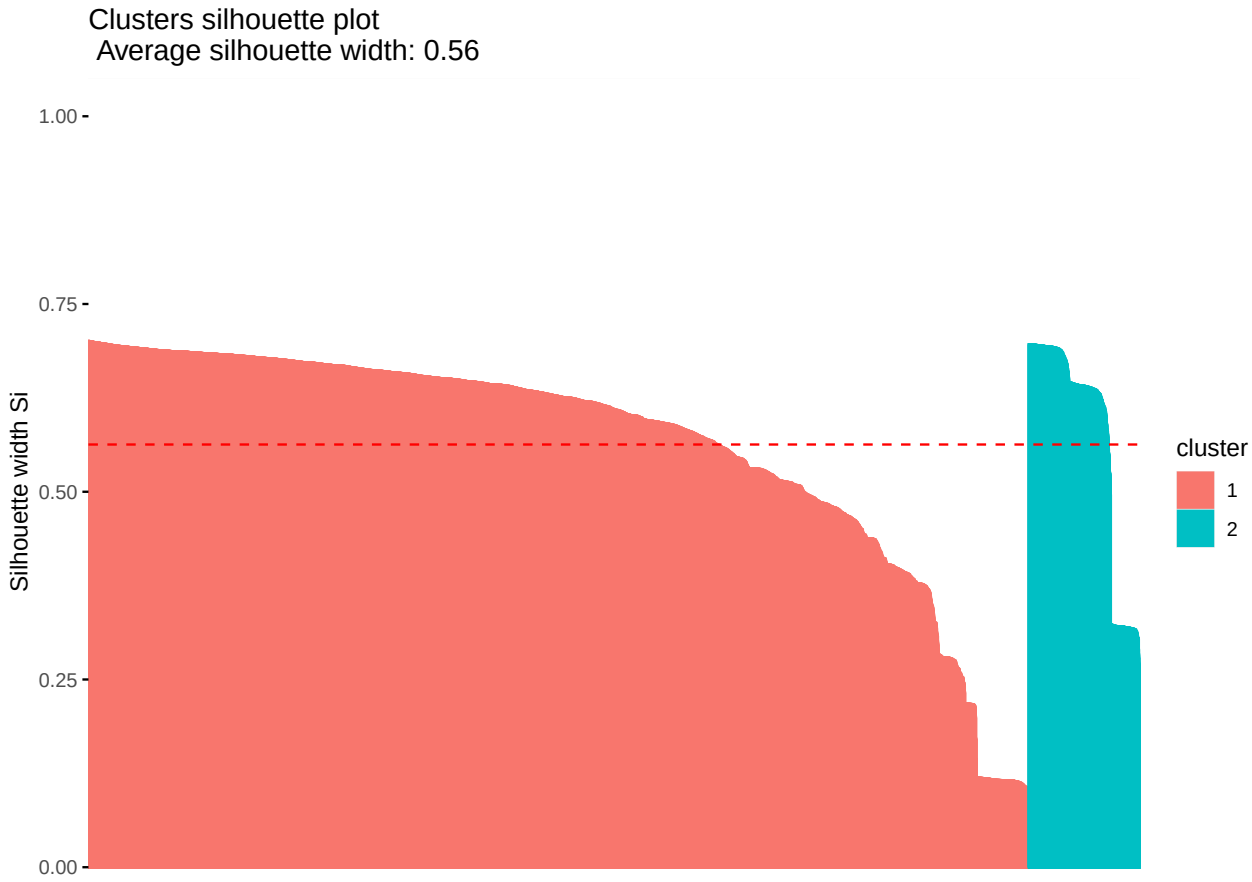
Health


Silhouette Coefficient
0.44
0.46
0.48
0.50
0.52
0.54
0.56
2
3
4
5
6
Number of Clusters


Cluster plot
Dim2 (20.2%)
Dim1 (36.5%)
2.5
0.0
−2.5
−5.0
0
5
10
209897
195433
243150
178661
378042
231325
181116
cluster
1
2



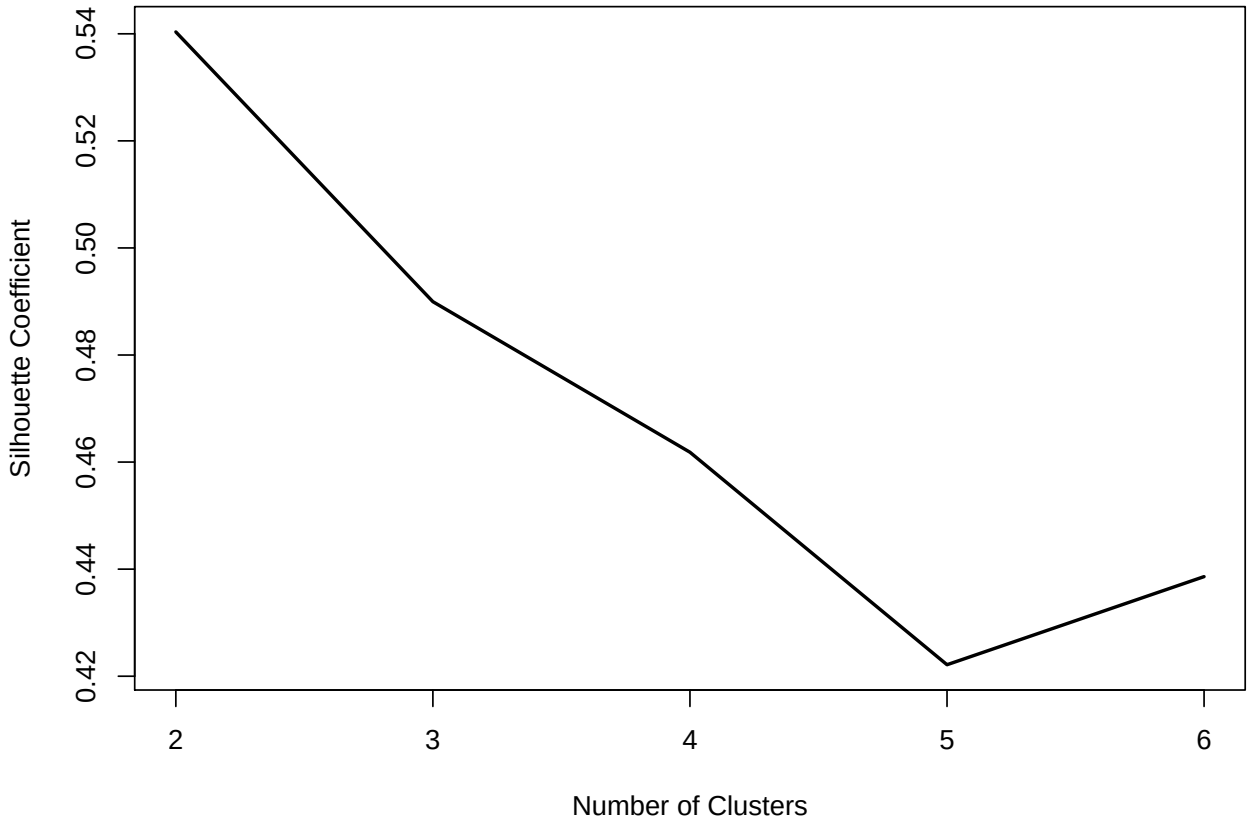

```
## [1] "Segment cutoff values:"
## [1] 0.009987632
##   cluster size ave.sil.width
## 1       1 7970          0.56
## 2       2  943          0.57
```



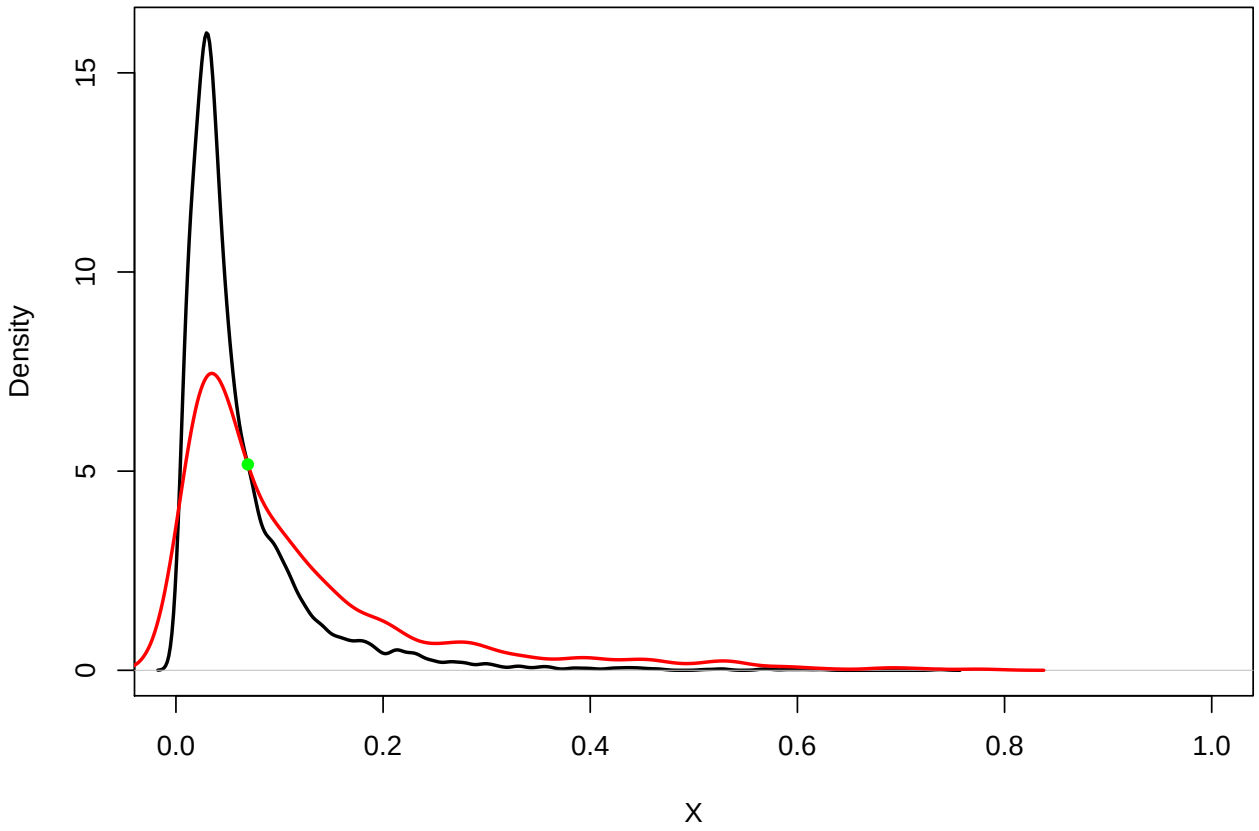

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##      13107      1583
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       71.7        79
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   234812.3  230156.2
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2
##        5718       720
## B      5474       641
## D      1452       166
## K       463        56
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2
##      0.228     0.235
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2
## Alberta                     412        46
## BritishColumbia             648        75
## NewBrunswick                115         7
## NorthwestTerritories          7         0
## NovaScotia                  403        47
## Ontario                    2007       233
## Quebec                      759        99
## Saskatchewan                 58        13
## NA's                       8698      1063
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0     11877      1417
## 1       949       135
## 2       133        14
## F       148        17
```

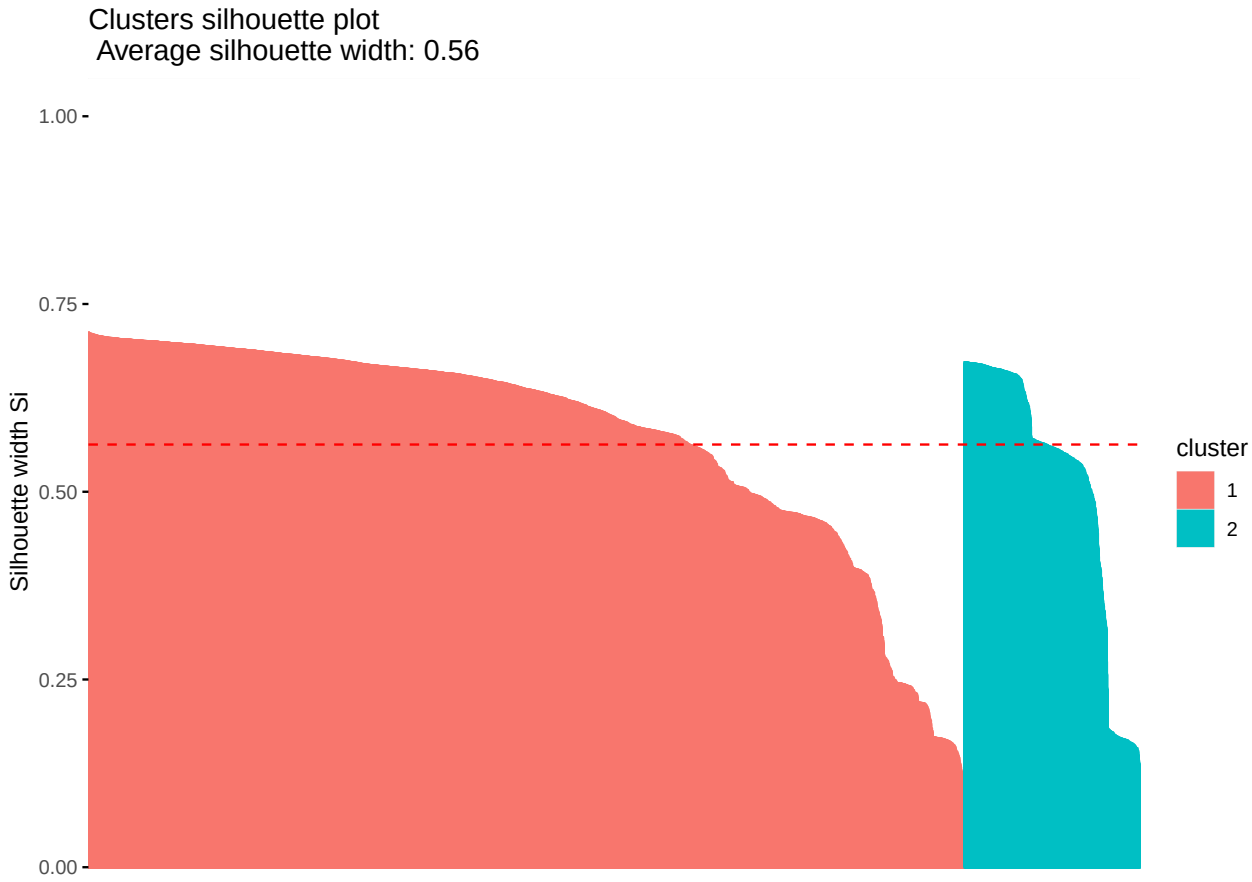
Grocery


Silhouette Coefficient
0.42
0.44
0.46
0.48
0.50
0.52
0.54
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



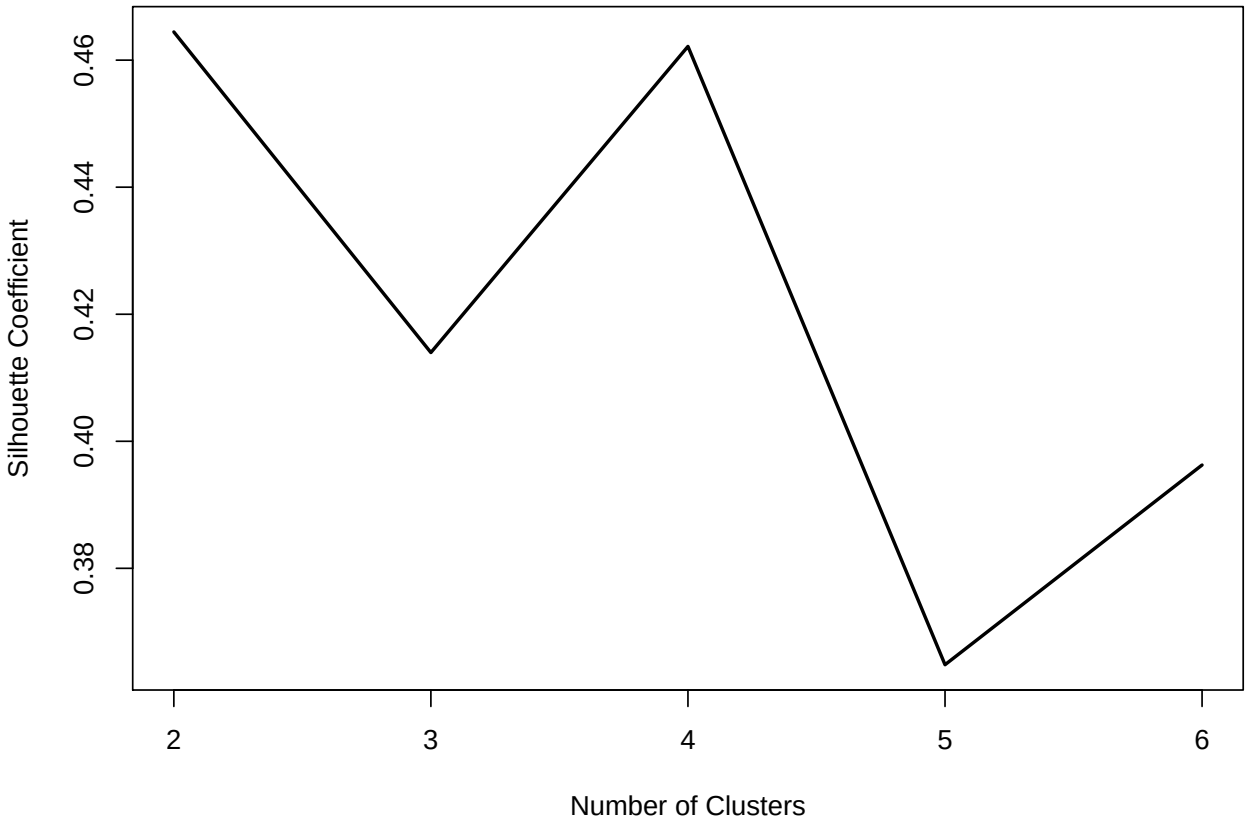

```
## [1] "Segment cutoff values:"
## [1] 0.06941786
##   cluster size ave.sil.width
## 1       1 3495          0.57
## 2       2  700          0.51
```

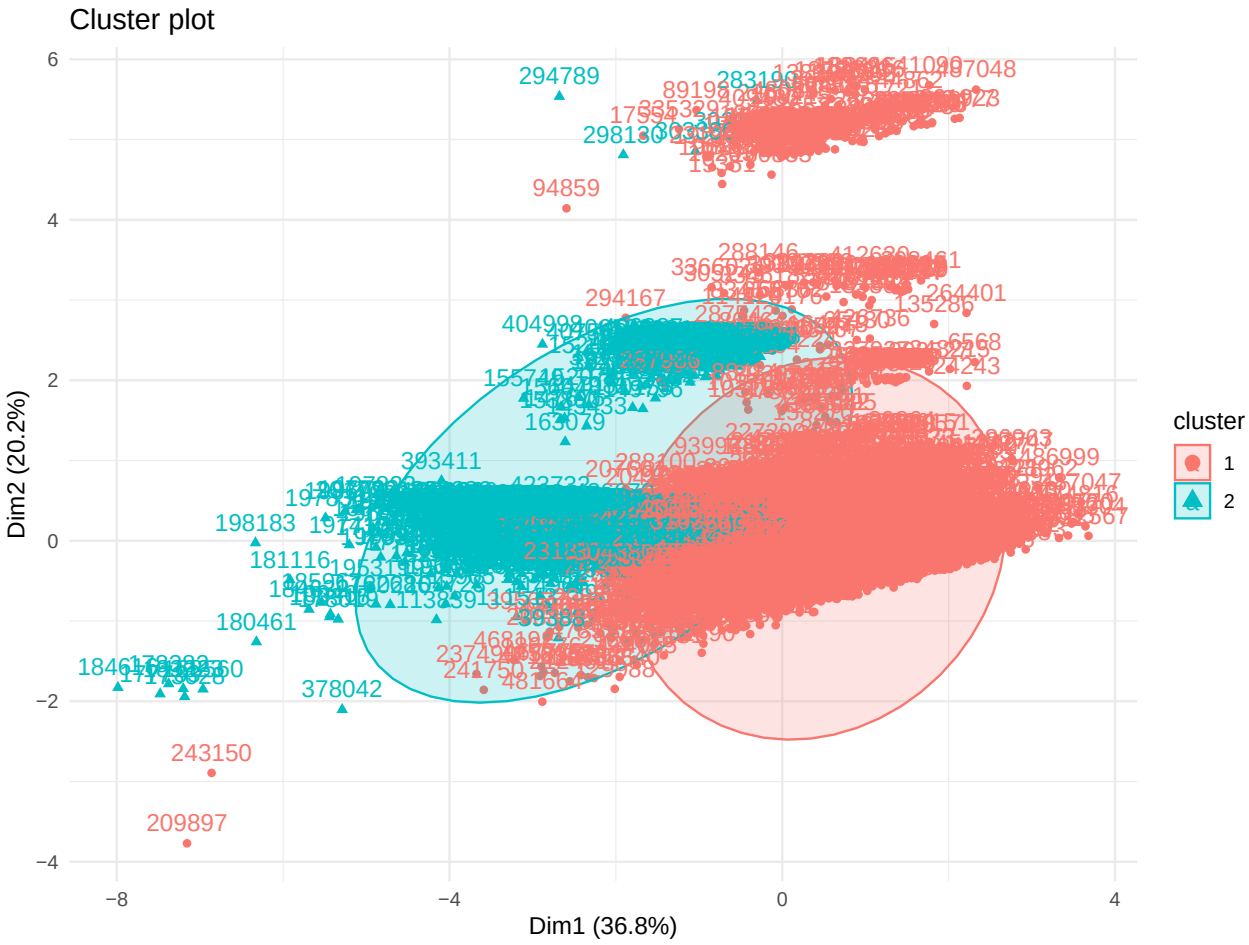


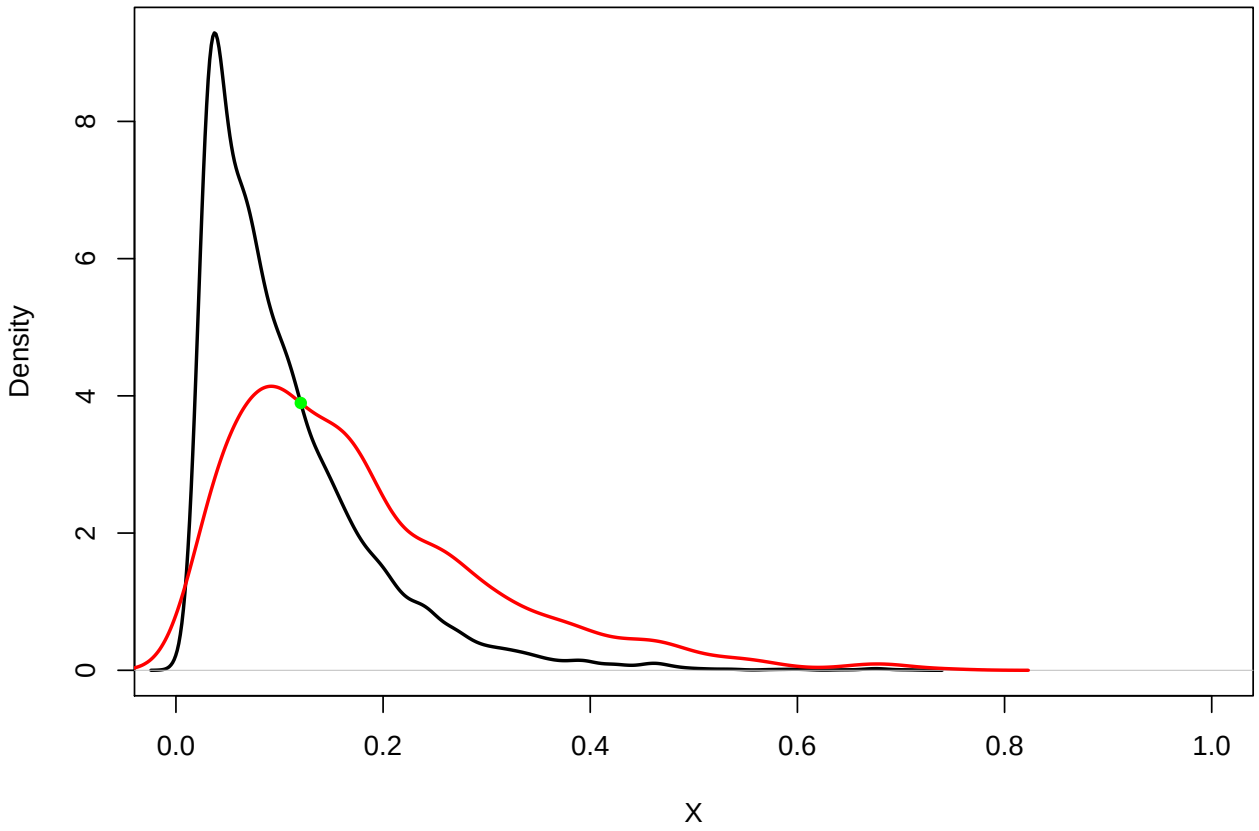

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##      12218      2472
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       72.5      72.6
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   232517.5  243173.9
##
##
##
##  CMA Type:
##    Cluster 1 Cluster 2
##         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
```

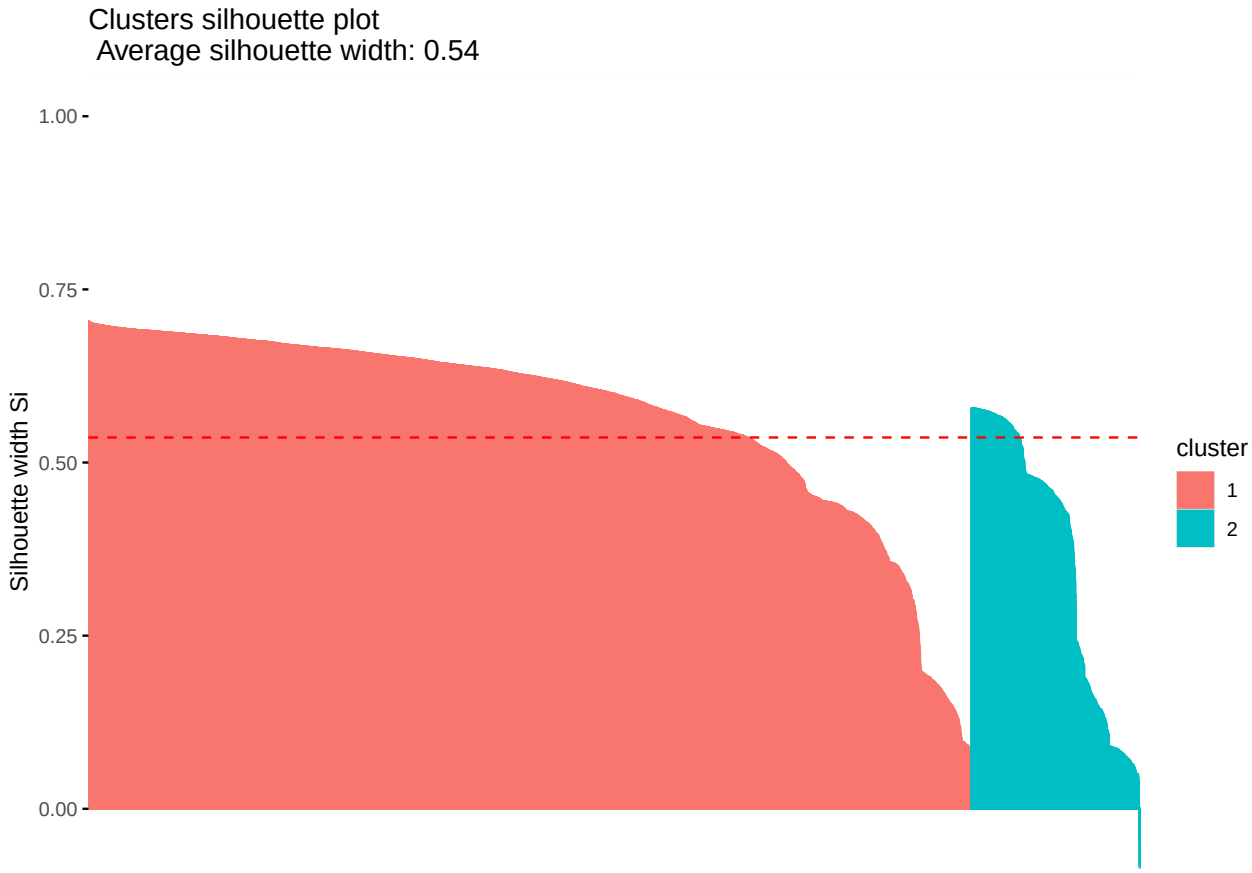
Primary Education




Cluster plot
Dim2 (20.2%)
Dim1 (36.8%)
cluster
1
2
294789
94859
294167
404998
163079
393411
198183
181116
180461
378042
243150
209897
6
4
2
0
−2
−4
−8



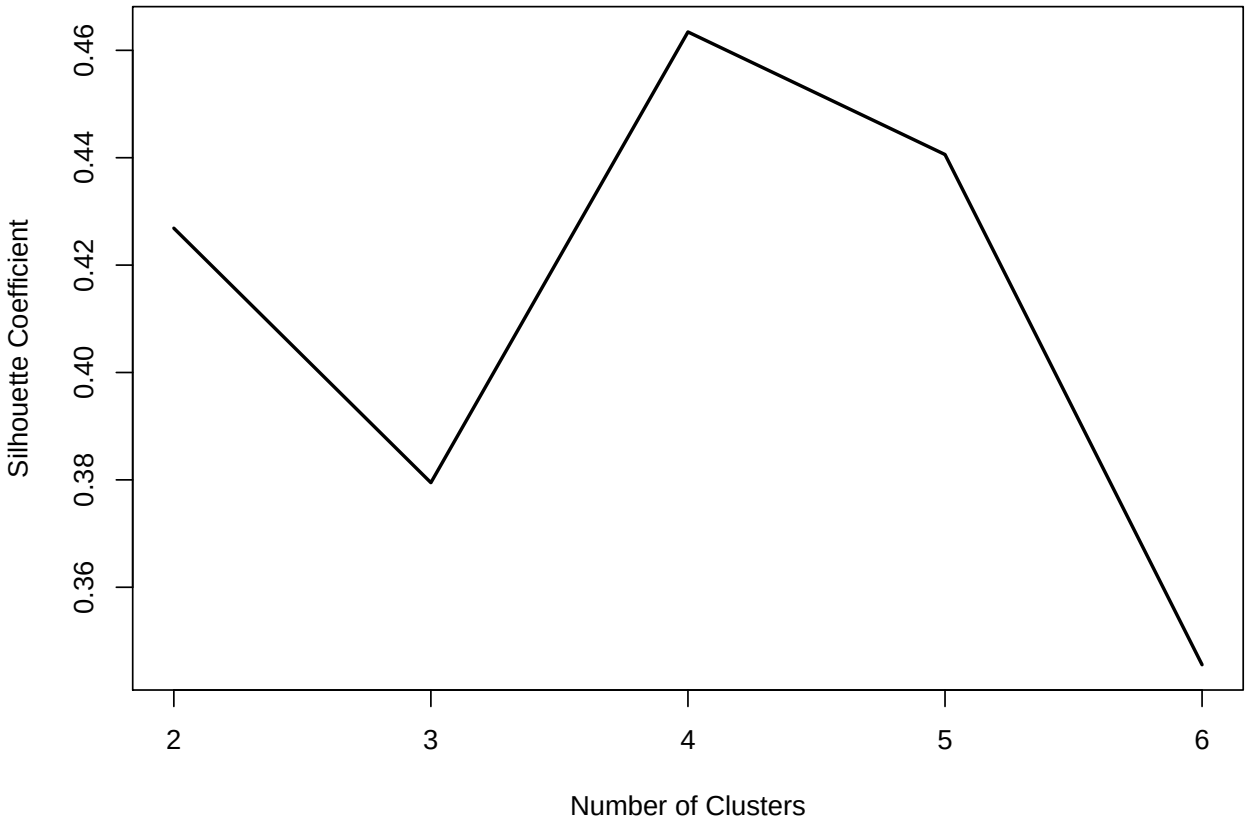

```
## [1] "Segment cutoff values:"
## [1] 0.1205326
##   cluster size ave.sil.width
## 1       1 5620          0.57
## 2       2 1071          0.36
```

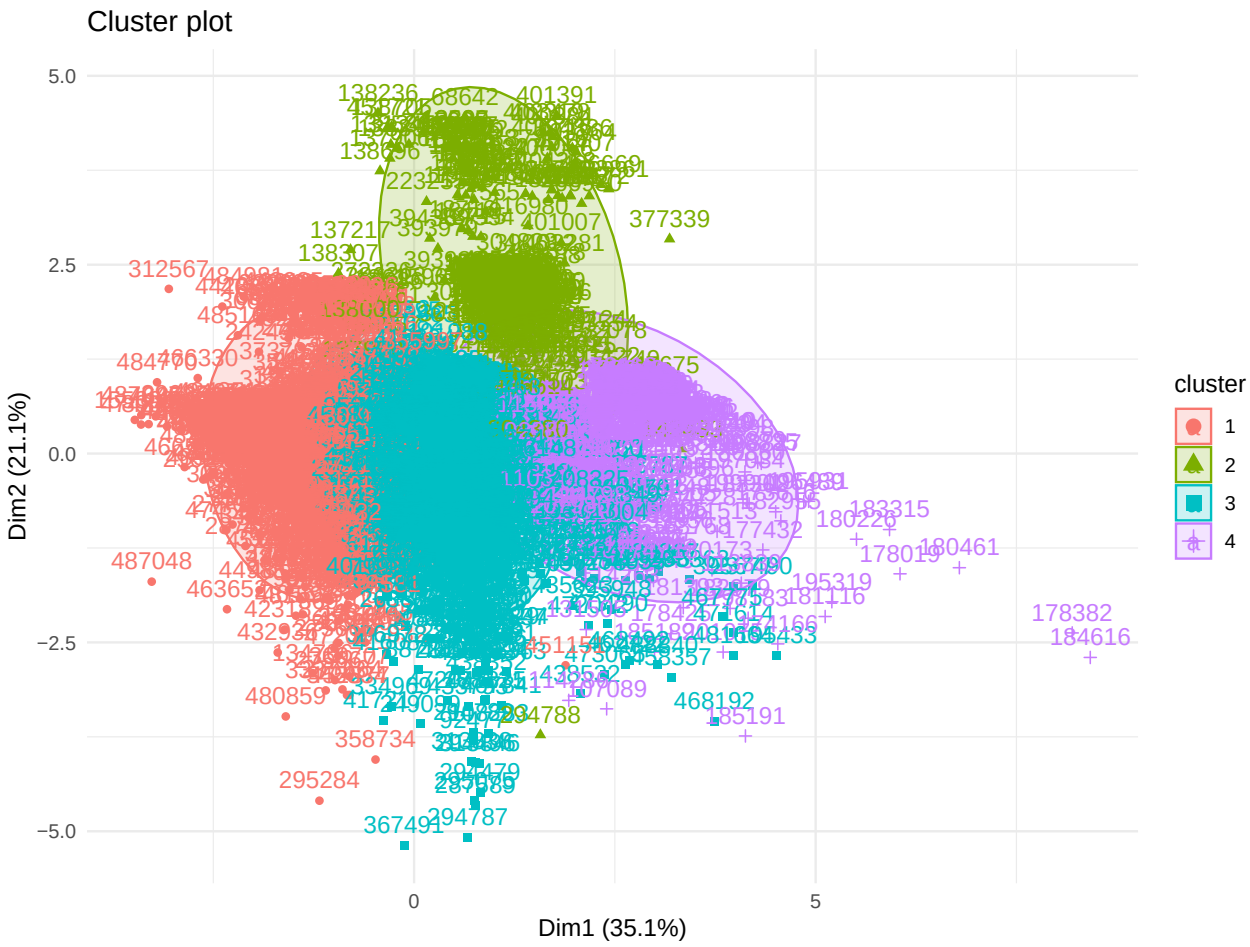


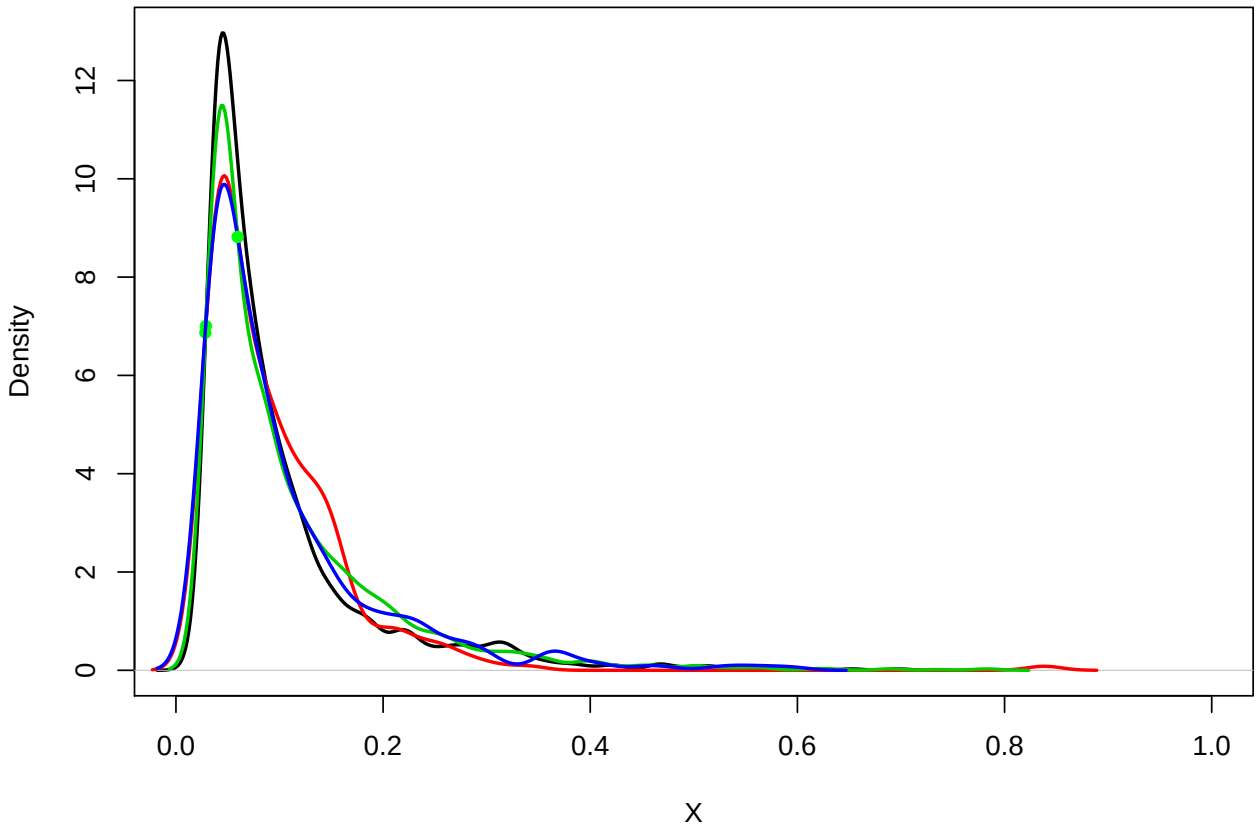

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##      12339      2351
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       73.1      69.5
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   237615.6  216973.5
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2
##        5380      1058
## B      5163       952
## D      1346       272
## K       450        69
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2
##      0.228      0.23
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2
## Alberta                     384        74
## BritishColumbia             596       127
## NewBrunswick                108        14
## NorthwestTerritories          6         1
## NovaScotia                  385        65
## Ontario                    1886       354
## Quebec                      733       125
## Saskatchewan                 60        11
## NA's                       8181      1580
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0     11176      2118
## 1       898       186
## 2       128        19
## F       137        28
```

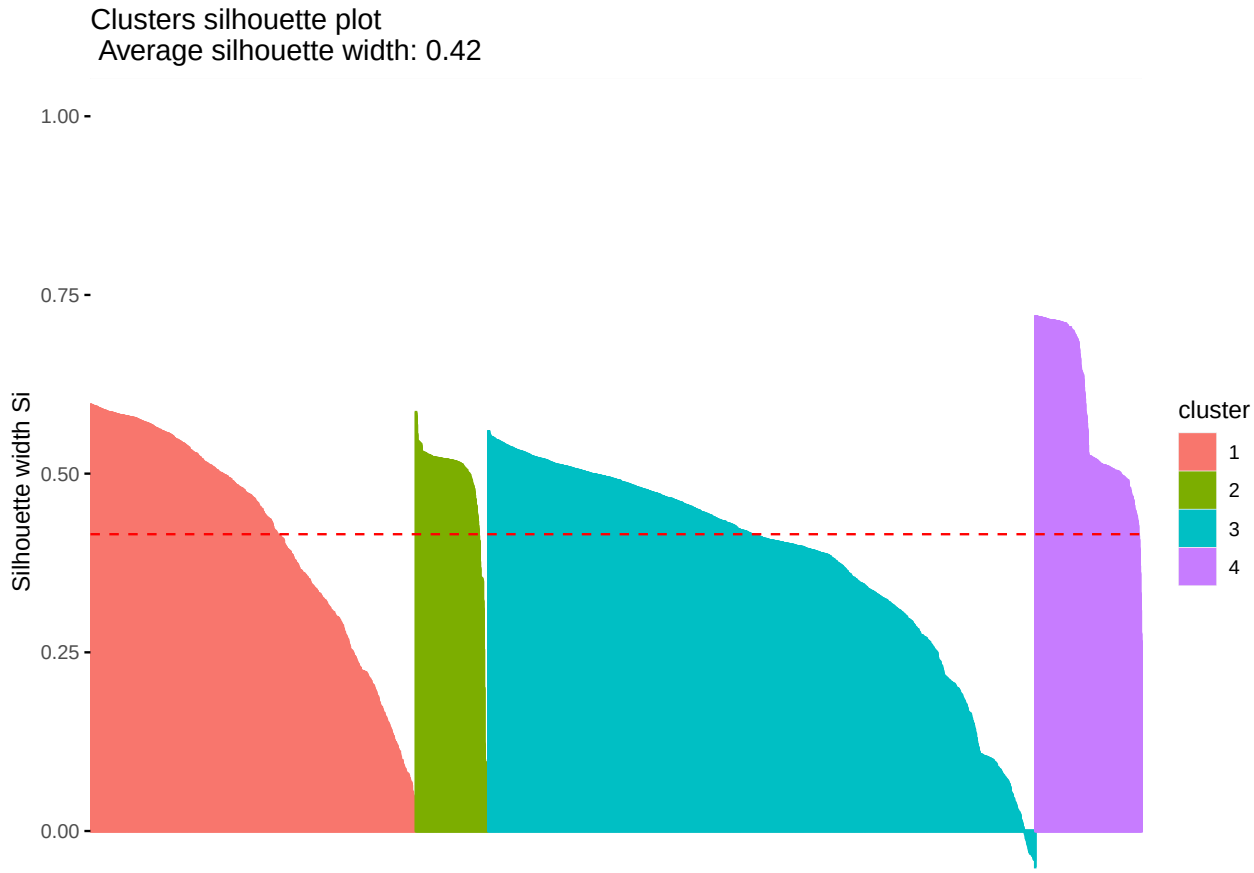
Secondary Education




Cluster plot
Dim2 (21.1%)
Dim1 (35.1%)
5.0
2.5
0.0
−2.5
−5.0
0
5
cluster
1
2
3
4



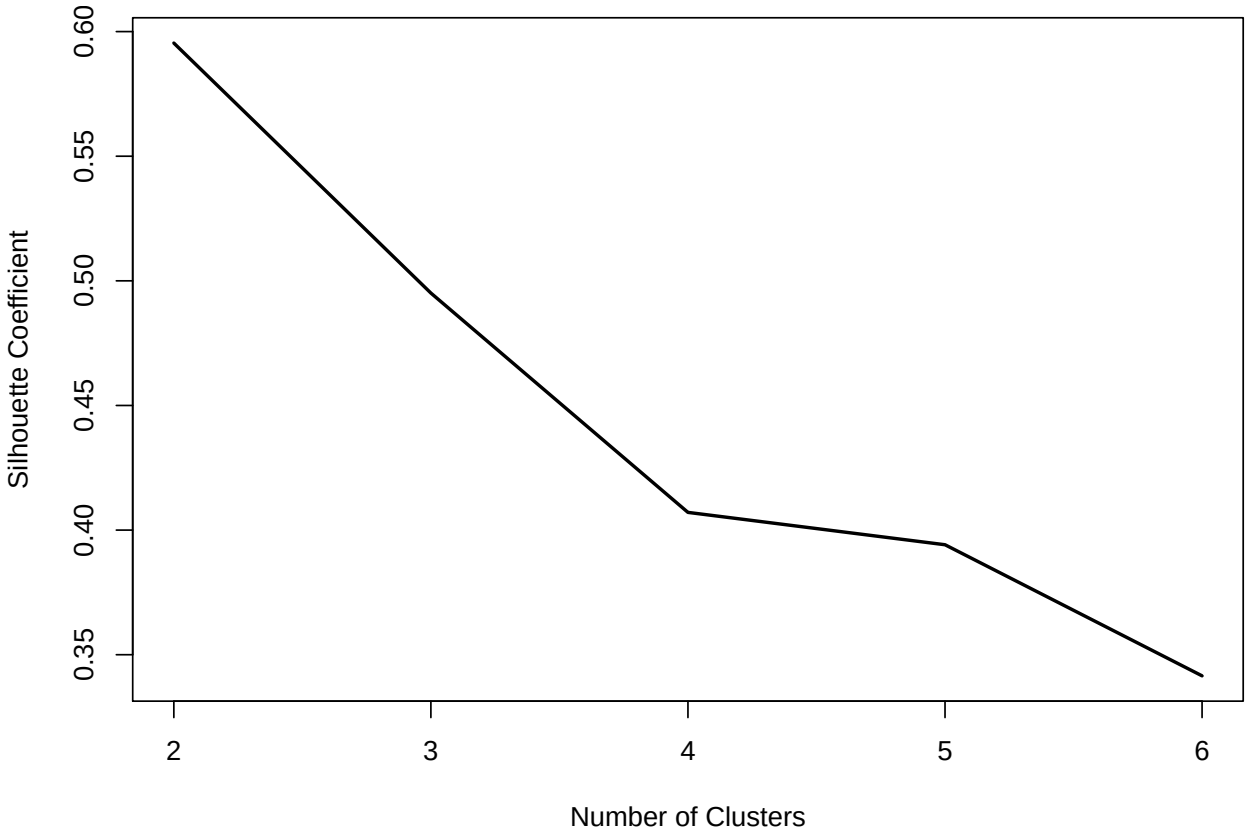

```
## [1] "Segment cutoff values:"
## [1] 0.02890393
## [1] 0.02846262
## [1] 0.05957614
##   cluster size ave.sil.width
## 1       1 1296          0.41
## 2       2  289          0.48
## 3       3 2179          0.37
## 4       4  425          0.59
```

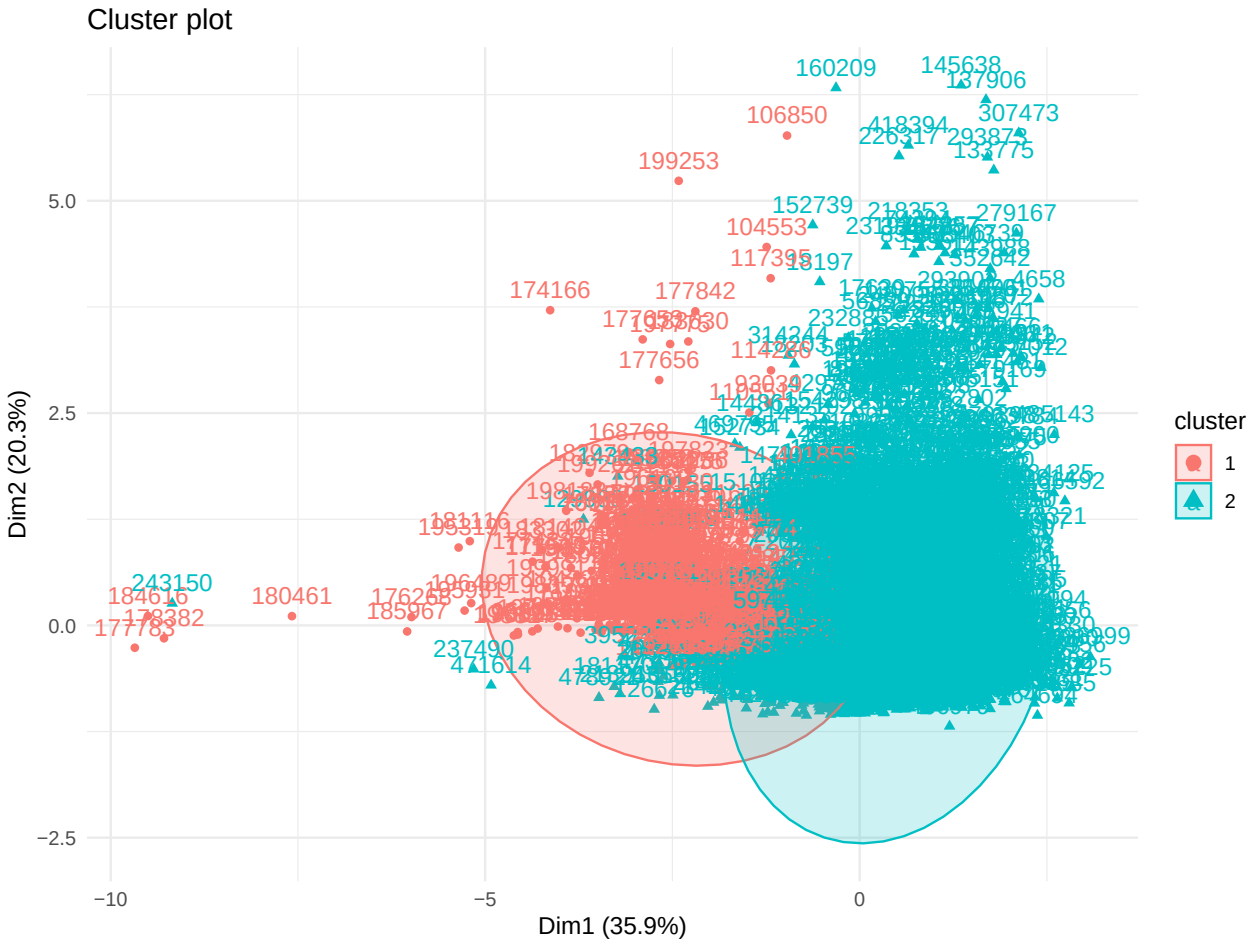


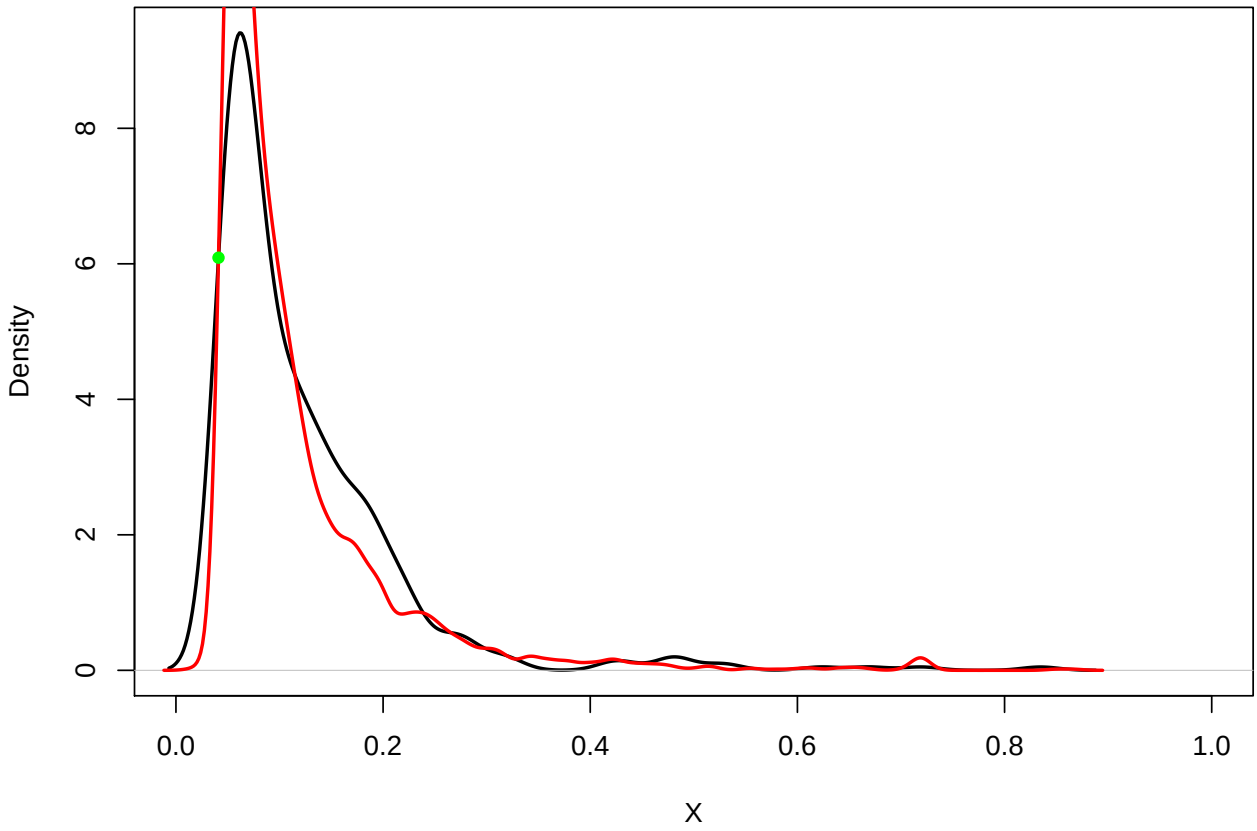

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##       4522      1009      7652      1507
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##       73.7      66.2      71.6      77.8
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##   228671.6  227297.5  240950.9  222207.6
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2 Cluster 3 Cluster 4
##        1984       426      3340       688
## B      1869       424      3213       609
## D       516       123       820       159
## K       153        36       279        51
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2 Cluster 3 Cluster 4
##      0.227     0.225     0.229     0.233
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2 Cluster 3 Cluster 4
## Alberta                     154        18       241        45
## BritishColumbia             225        53       366        79
## NewBrunswick                 40         6        66        10
## NorthwestTerritories          4         1         2         0
## NovaScotia                  126        36       253        35
## Ontario                     663       177      1172       228
## Quebec                      267        59       456        76
## Saskatchewan                 16         3        46         6
## NA's                       3027       656      5050      1028
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2 Cluster 3 Cluster 4
## 0      4083       910      6943      1358
## 1       334        76       562       112
## 2        49        10        70        18
## F        56        13        77        19
```

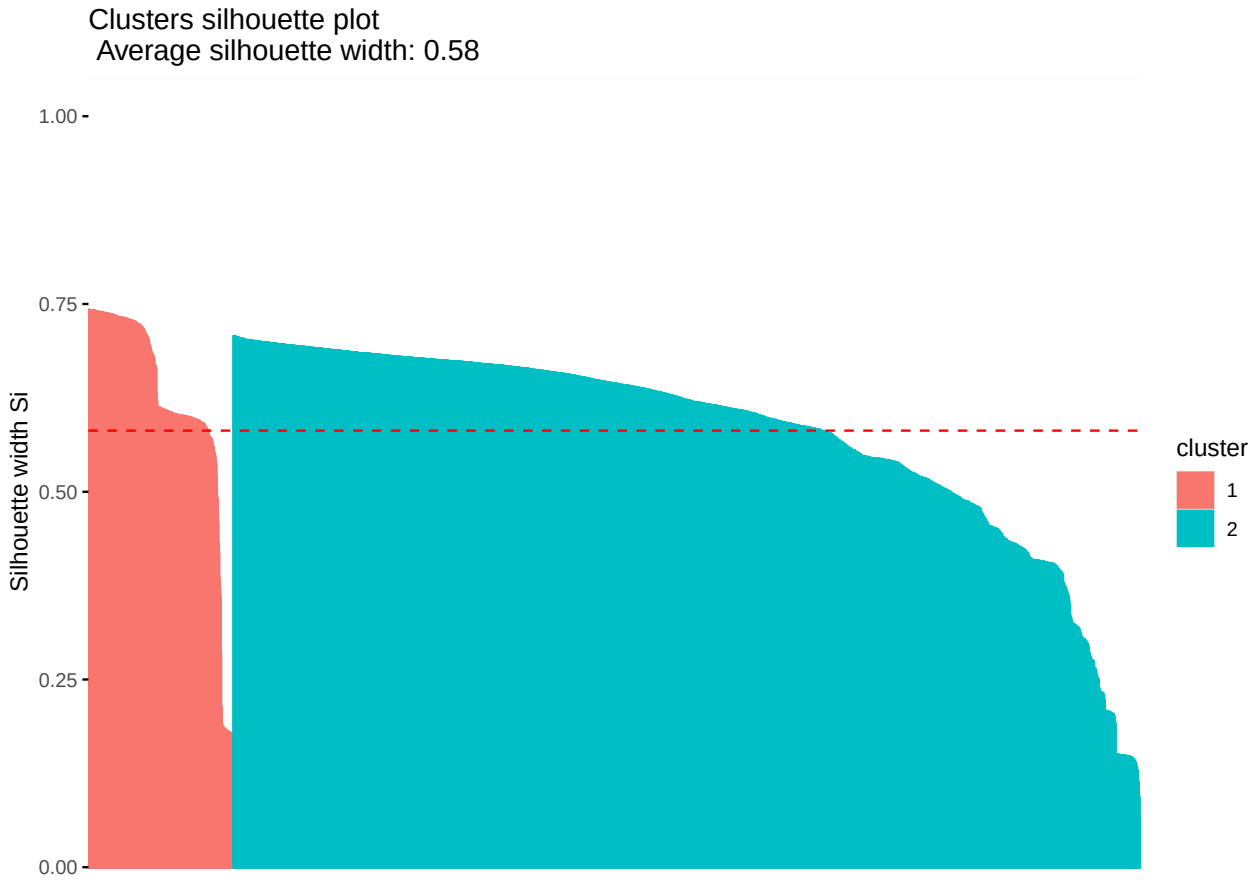
Libraries


Silhouette Coefficient
0.35
0.40
0.45
0.50
0.55
0.60
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



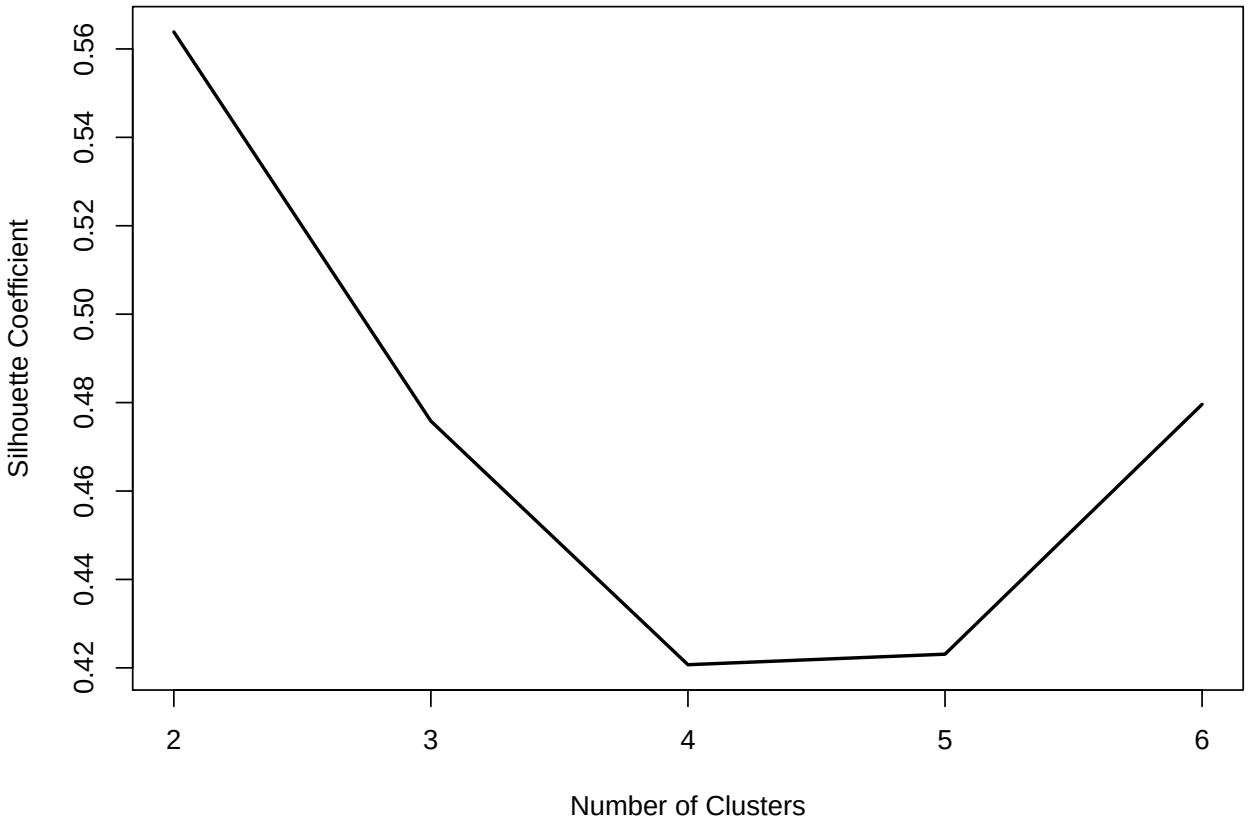

```
## [1] "Segment cutoff values:"
## [1] 0.04110552
##   cluster size ave.sil.width
## 1       1  454          0.61
## 2       2 2836          0.58
```

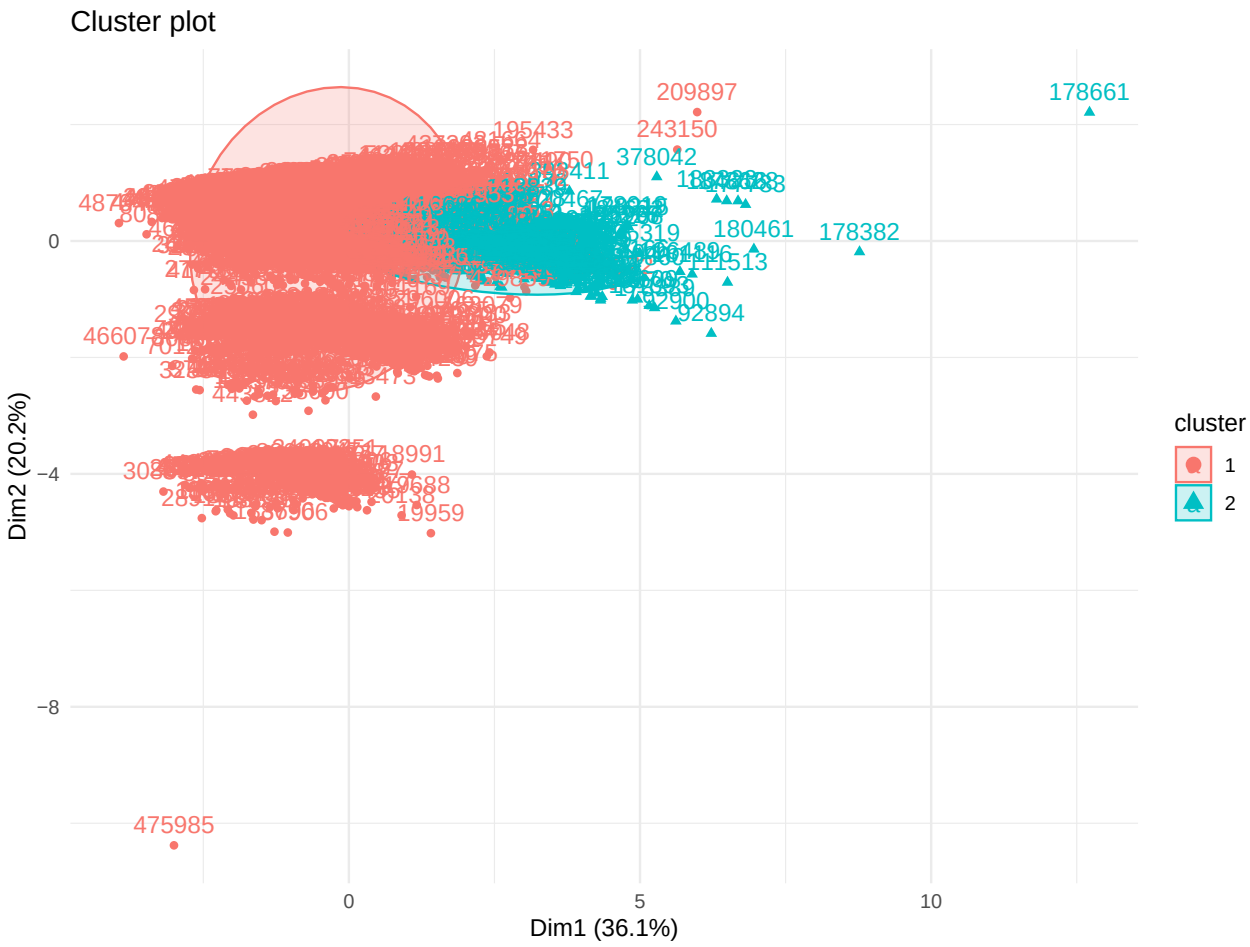


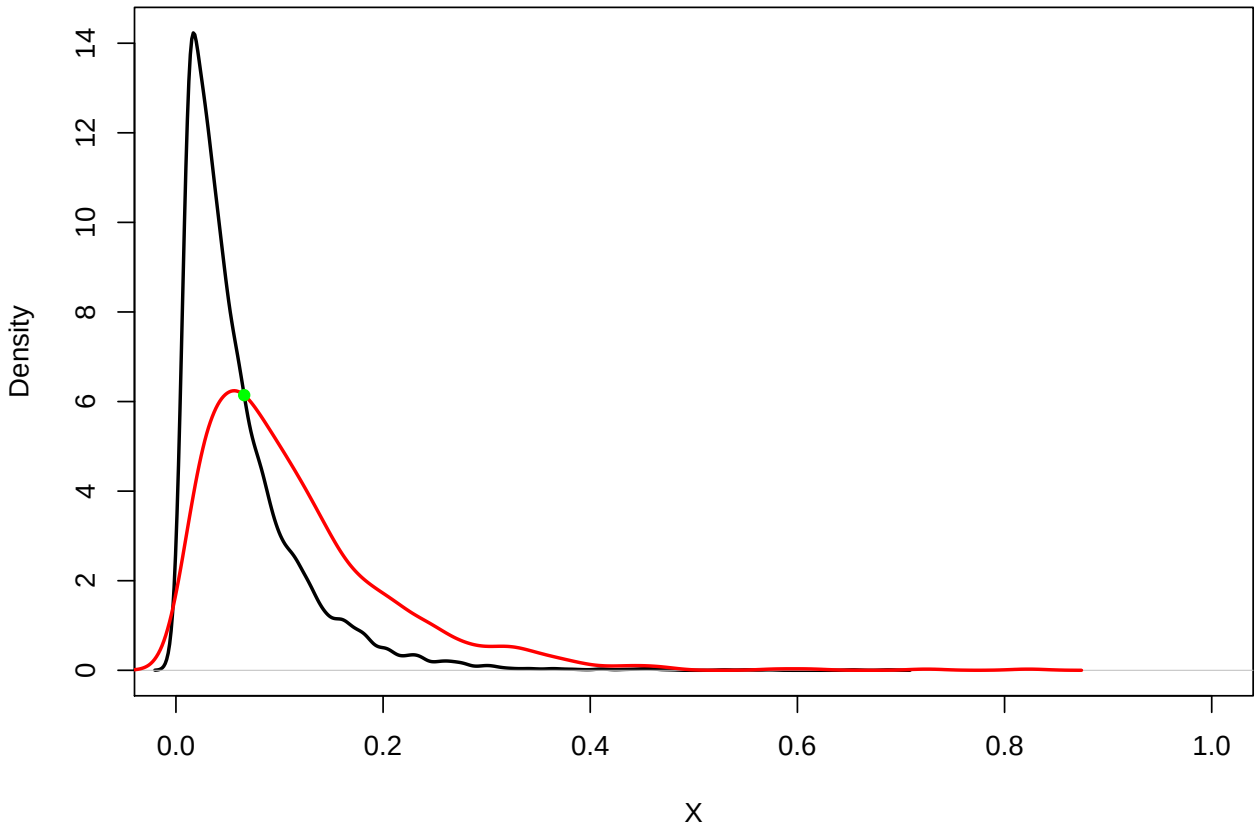

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##       2037     12653
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       71.2      72.7
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   232758.6  234560.4
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2
##         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
```

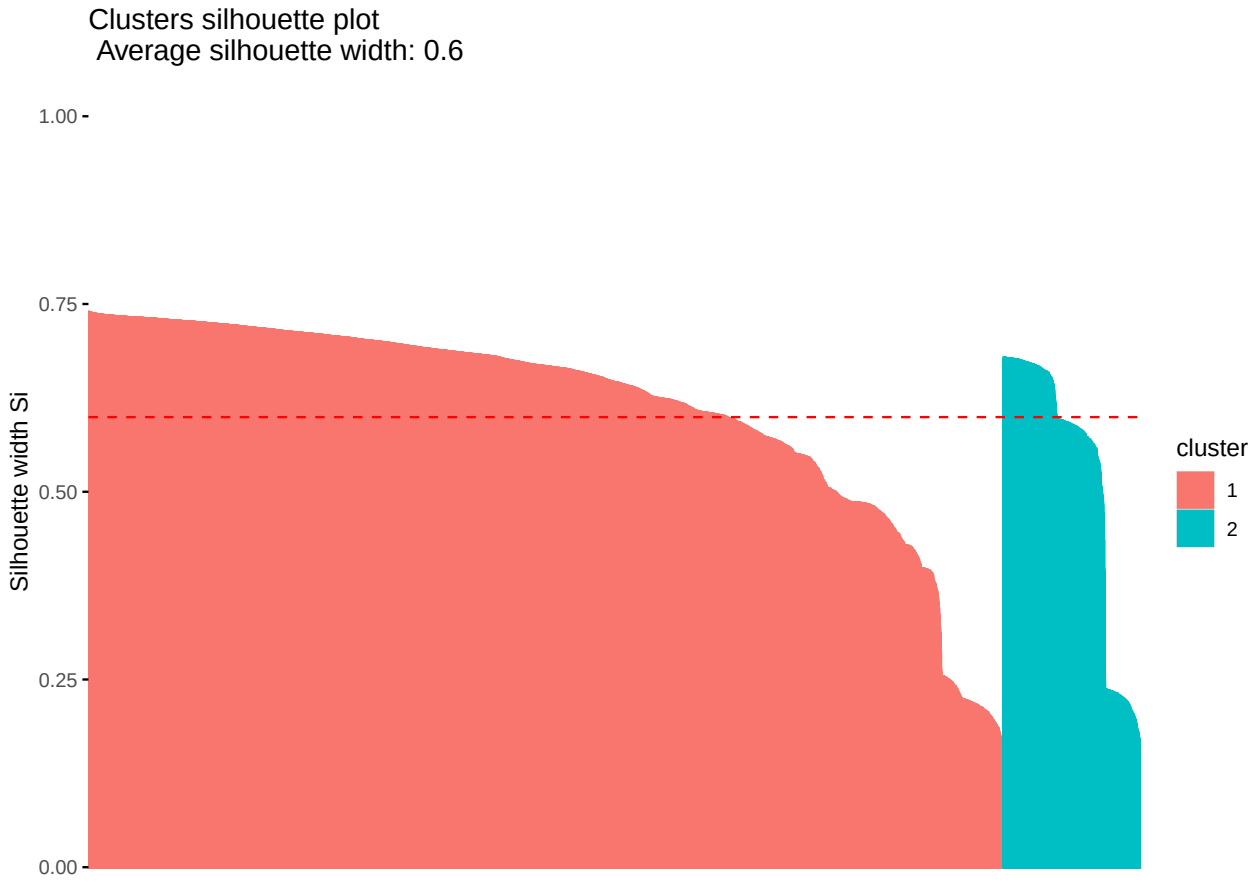
Parks


Silhouette Coefficient
0.42
0.44
0.46
0.48
0.50
0.52
0.54
0.56
2
3
4
5
6
Number of Clusters


Cluster plot
209897
178661
243150
195433
378042
180461
178382
92894
475985
Dim2 (20.2%)
Dim1 (36.1%)
0
−4
−8
0
5
10
cluster
1
2



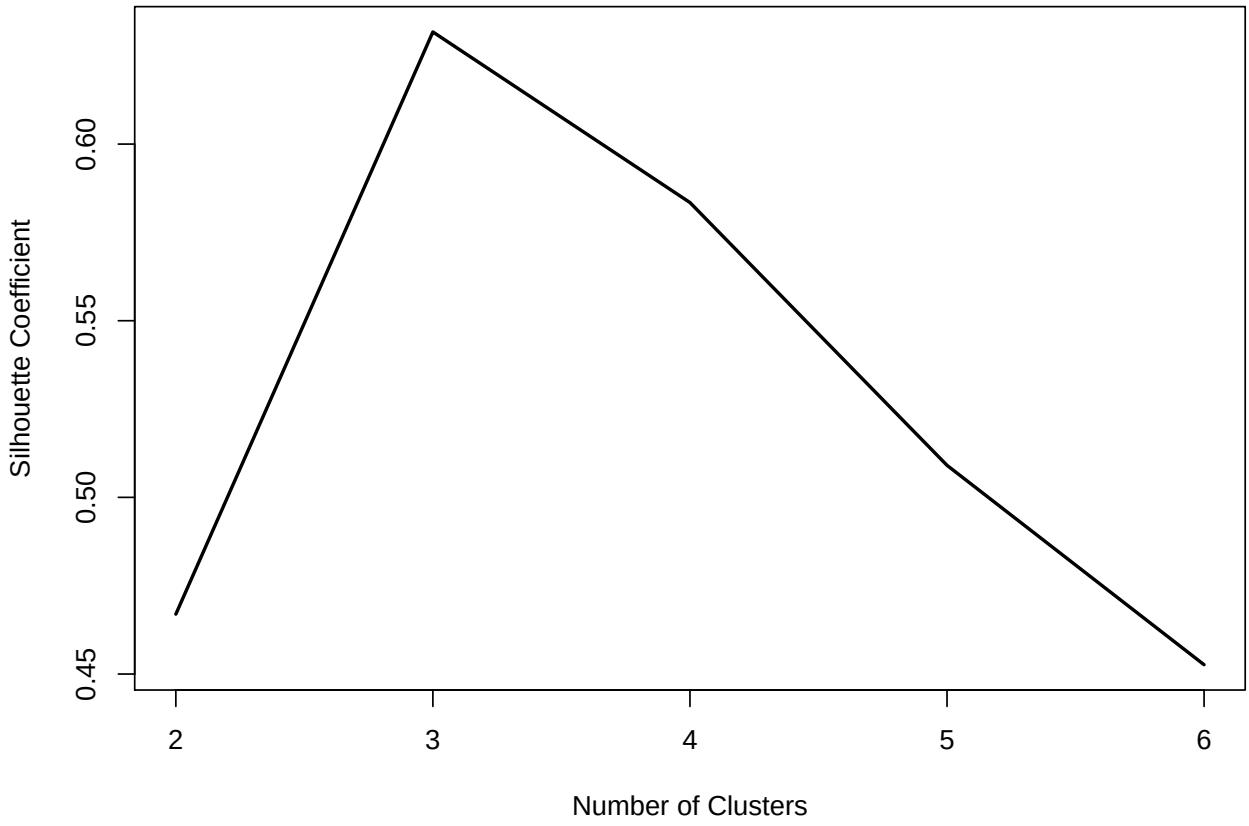

```
## [1] "Segment cutoff values:"
## [1] 0.0658873
##   cluster size ave.sil.width
## 1       1 6065          0.61
## 2       2  904          0.52
```

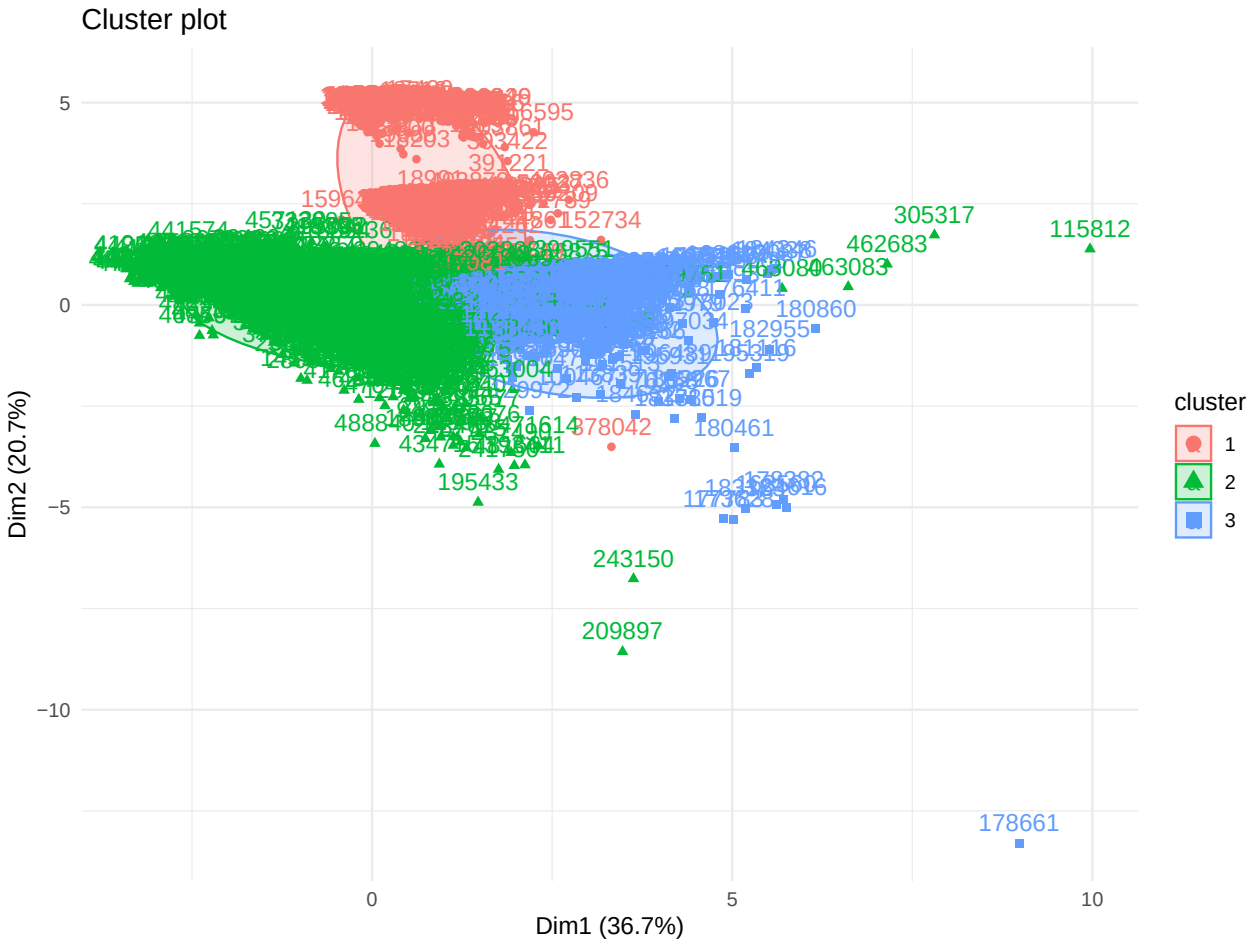


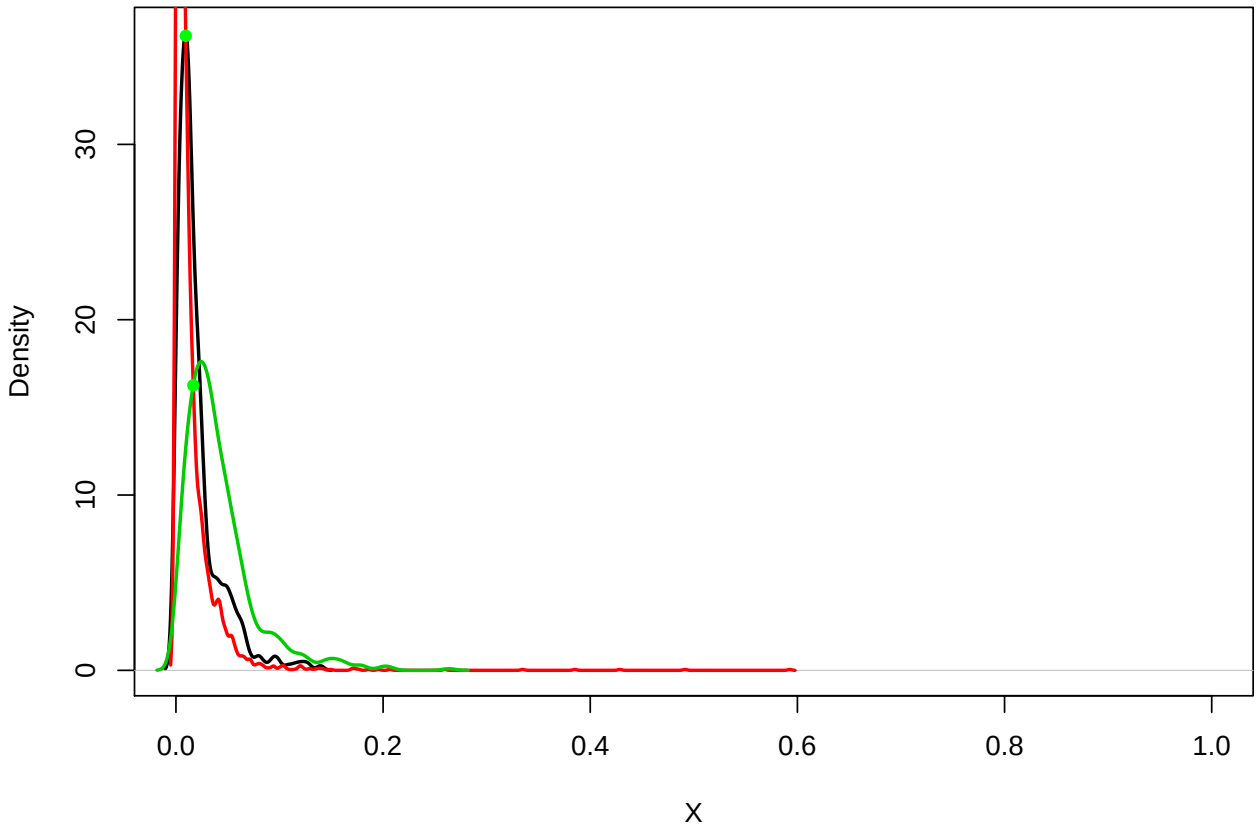

```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##      12788      1902
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       73.2      68.1
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   236951.3  216564.9
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2
##        5585       853
## B      5350       765
## D      1399       219
## K       454        65
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2
##      0.228     0.233
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2
## Alberta                     401        57
## BritishColumbia             628        95
## NewBrunswick                107        15
## NorthwestTerritories          7         0
## NovaScotia                  385        65
## Ontario                    1948       292
## Quebec                      756       102
## Saskatchewan                 62         9
## NA's                       8494      1267
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0     11566      1728
## 1       955       129
## 2       128        19
## F       139        26
```

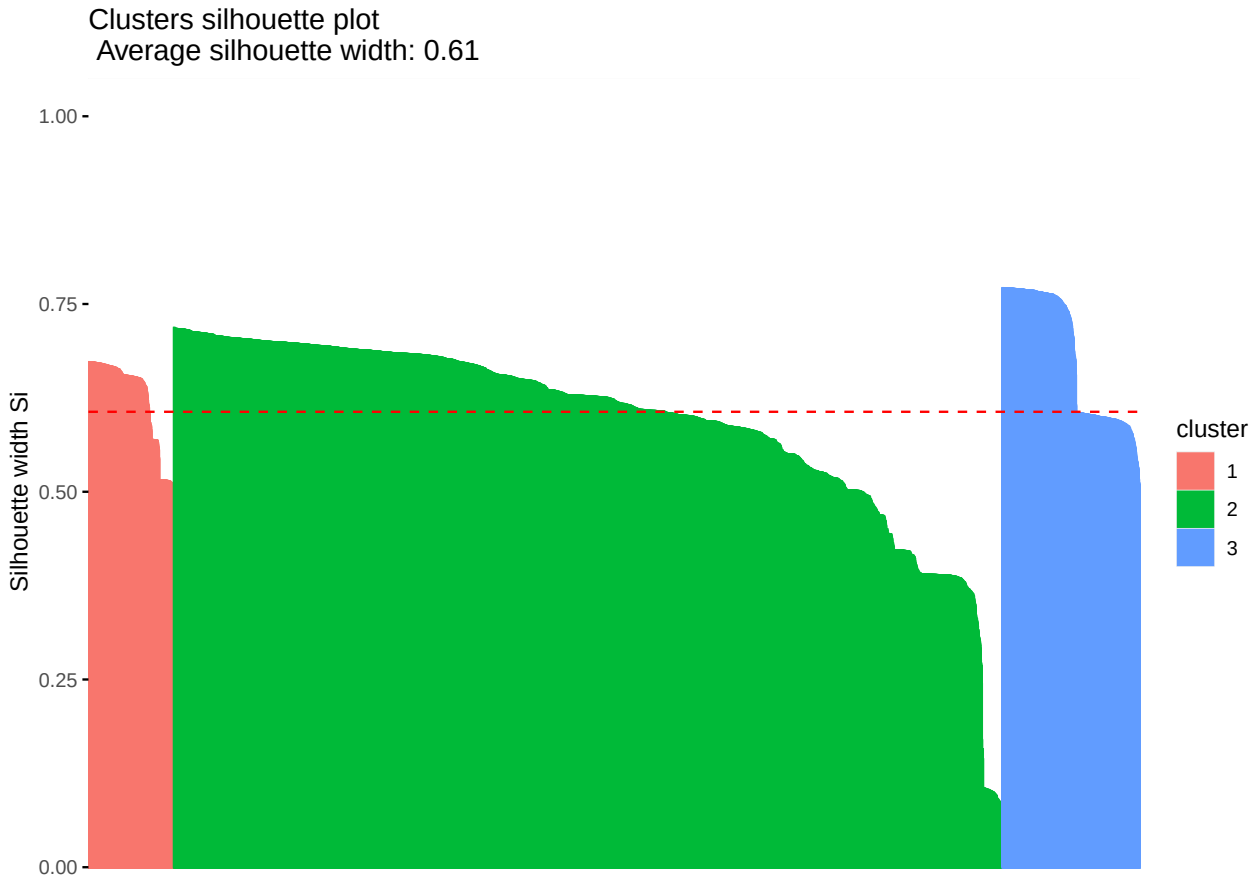
Transit


Silhouette Coefficient
0.45
0.50
0.55
0.60
2
3
4
5
6
Number of Clusters


Cluster plot
Dim2 (20.7%)
Dim1 (36.7%)
cluster
1
2
3
5
0
−5
−10
0
5
10
305317
115812
462683
180860
182955
180461
878042
195433
243150
209897
178661




```
## [1] "Segment cutoff values:"
## [1] 0.009520392
## [1] 0.01673738
##   cluster size ave.sil.width
## 1       1  435          0.62
## 2       2 4217          0.59
## 3       3  700          0.68
```




```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2 Cluster 3
##       1200     11559      1931
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2 Cluster 3
##         75      72.7      69.8
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2 Cluster 3
##   242255.8  233104.2  236596.6
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2 Cluster 3
##         540      5038       860
## B       482      4820       813
## D       140      1273       205
## K        38       428        53
```

```
##
##
##
##  Index of Remoteness:
##  Cluster 1 Cluster 2 Cluster 3
##       0.23     0.228     0.229
##
##
##
##  Provinces:
##                       Cluster 1 Cluster 2 Cluster 3
## Alberta                      34       372        52
## BritishColumbia              56       562       105
## NewBrunswick                 10        97        15
## NorthwestTerritories          0         7         0
## NovaScotia                   40       349        61
## Ontario                     163      1771       306
## Quebec                       81       671       106
## Saskatchewan                  7        57         7
## NA's                        809      7673      1279
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2 Cluster 3
## 0      1094     10432      1768
## 1        77       880       127
## 2        16       115        16
## F        13       132        20
```

Conclusion

text