National Institute Of Technology – Calicut
Data Mining

R programming

T.G. Deshan K. Sumanathilaka
B150413CS
Computer Science and Engineering
(B.Tech)

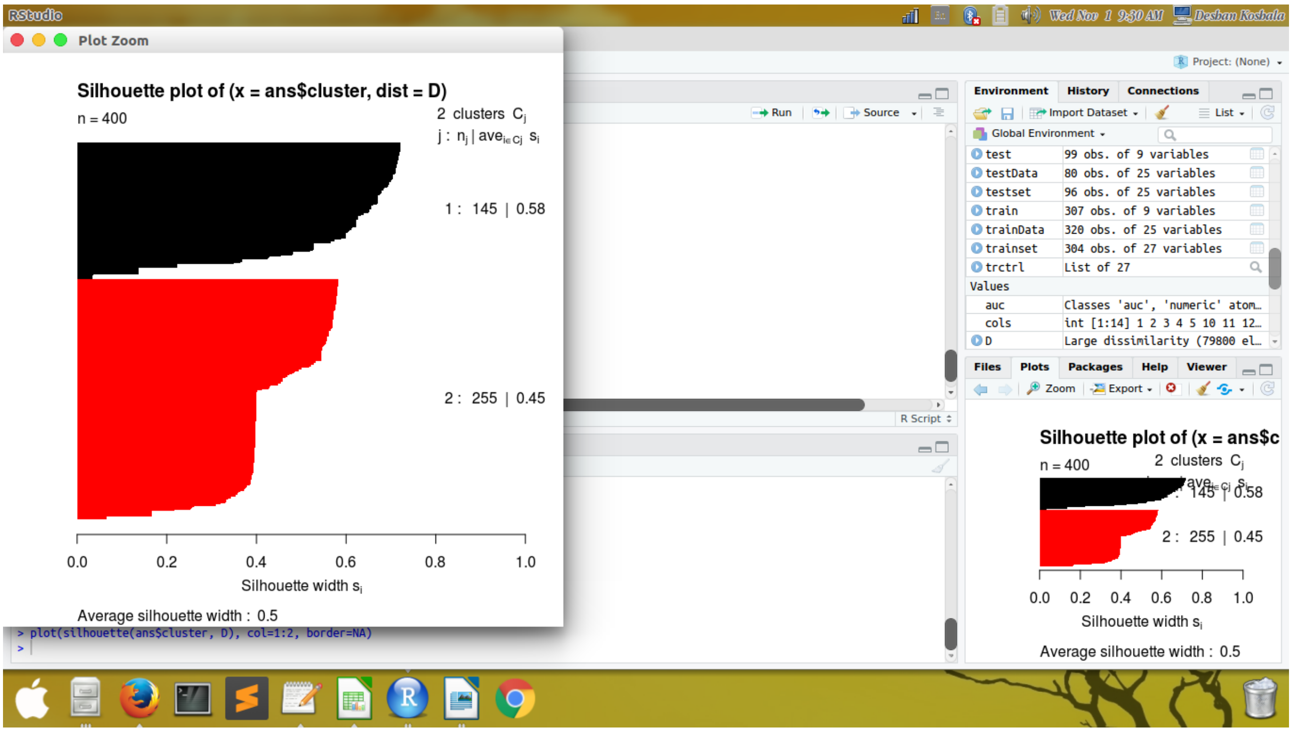
DATA-MINING ASSIGNMENT #02

A) K means clustering with k=actual number of classes

```
kidney_noc <- read.csv("/home/deshan/Desktop/clean_kidney_kmean.csv")
cols <- c(1:5, 10:18)
ans <- kmeans(kidney_noc[, cols], 2)
ans

K-means clustering with 2 clusters of sizes 145, 255

Cluster means:
          age          bp          sg           al           su         bgr           bu
sc         sod         pot        hemo
1 48.05850604 75.09936011 1.019482759 0.7379310345 0.1862068966 135.4104223 55.08658160
2.961808759 137.0499835 4.759555779 13.09259612
2 53.43085101 77.24792803 1.016705882 0.9921568627 0.5137254902 155.2160608 58.75582112
3.135370384 137.8009961 4.552007388 12.20450309
          pcv        wbcc        rbcc
1 40.46657582 6227.586207 4.771752345
2 37.98488574 9644.897959 4.670862333

Clustering vector:
  [1] 1 1 1 1 1 1 2 1 2 2 2 1 2 2 2 1 2 2 2 1 2 1 1 2 2 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2 2
1 2 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 1 2 2 2
 [67] 2 2 2 1 2 2 2 1 1 2 1 1 2 2 2 1 2 2 1 2 2 1 2 2 2 2 1 1 2 2 2 2 1 2 2 2 1 2 2 1 1 2 2
2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 1 2 2 2 1 2 2 2
[133] 2 2 2 2 2 2 2 2 2 1 2 2 2 1 2 2 2 2 1 2 2 2 2 1 2 1 2 2 2 2 1 2 2 2 2 1 1 2 2 2 2 2 2
1 2 2 2 2 2 2 2 2 1 2 2 2 2 1 2 2 2 1 2 2 2 1
[199] 2 2 1 2 2 2 2 2 2 2 1 1 2 2 2 1 1 2 2 2 1 2 2 2 2 2 2 2 1 1 2 2 2 2 2 2 1 1 2 2 2 2 1
1 2 2 2 1 1 2 2 1 2 2 2 1 2 2 2 2 2 1 2 1 1 2
[265] 1 2 1 2 2 1 1 2 1 2 2 1 1 1 2 1 2 2 1 2 1 1 2 2 2 1 2 1 2 1 2 2 1 1 2 2 2 1 2 1 1 2 1
1 1 2 2 1 1 2 2 1 1 2 1 2 2 2 1 1 2 1 2 2 2 1
[331] 2 1 2 1 1 2 2 1 1 1 1 1 1 1 2 1 1 1 1 2 2 2 1 1 1 1 1 2 1 1 1 1 2 1 2 2 2 1 2 2 2 1 2
2 2 2 2 1 1 1 1 1 1 1 2 2 1 1 2 1 1 1 1 1 2 1
[397] 1 1 1 1

Within cluster sum of squares by cluster:
[1]  199256539.1 1264972635.5
 (between_SS / total_SS =  42.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
"size"         "iter"
[9] "ifault"
```

** install.packages("cluster")

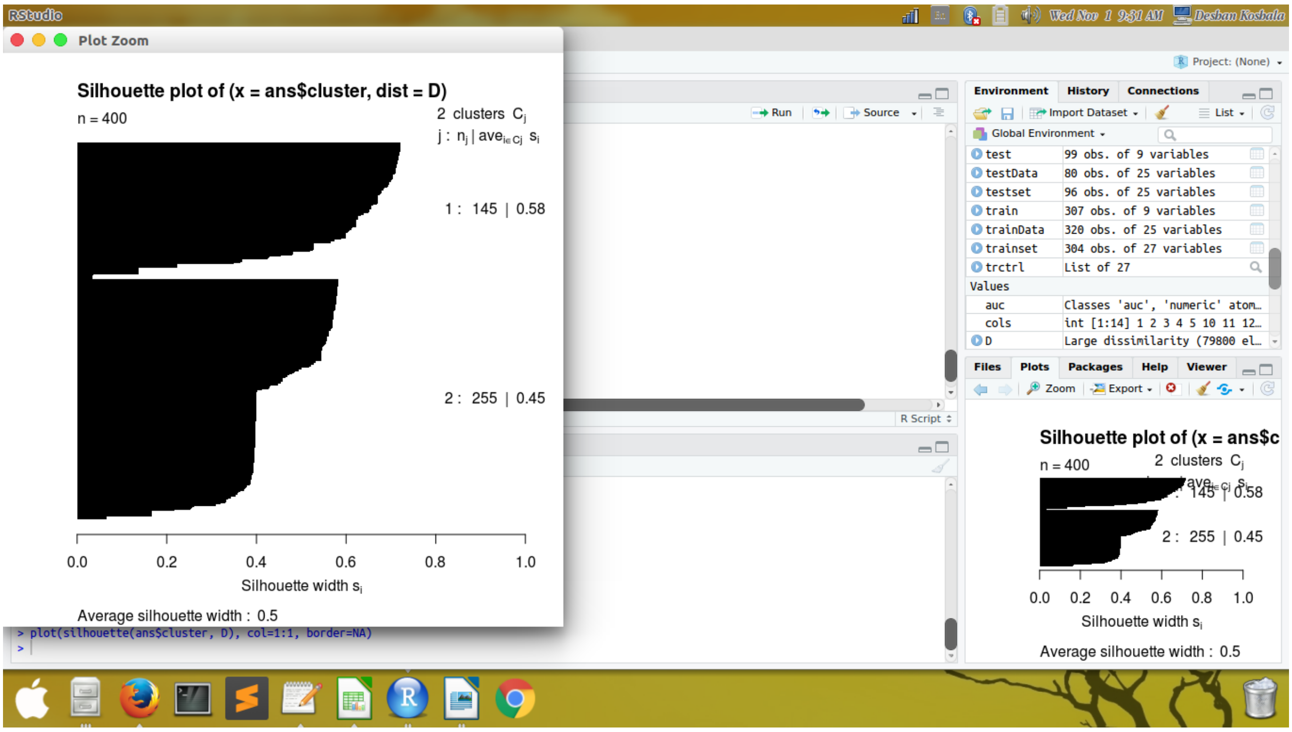
```
> library(cluster)
> D<- daisy(kidney_noc[, cols])

> plot(silhouette(ans$cluster, D), col=1:2, border=NA)
```



```
test        99 obs. of 9 variables
testData    80 obs. of 25 variables
testset     96 obs. of 25 variables
train       307 obs. of 9 variables
trainData   320 obs. of 25 variables
trainset    304 obs. of 27 variables
trctrl      List of 27
Values
  auc       Classes 'auc', 'numeric' atom…
  cols      int [1:14] 1 2 3 4 5 10 11 12…
D           Large dissimilarity (79800 el…
```

```
> plot(silhouette(ans$cluster, D), col=1:2, border=NA)
>
```

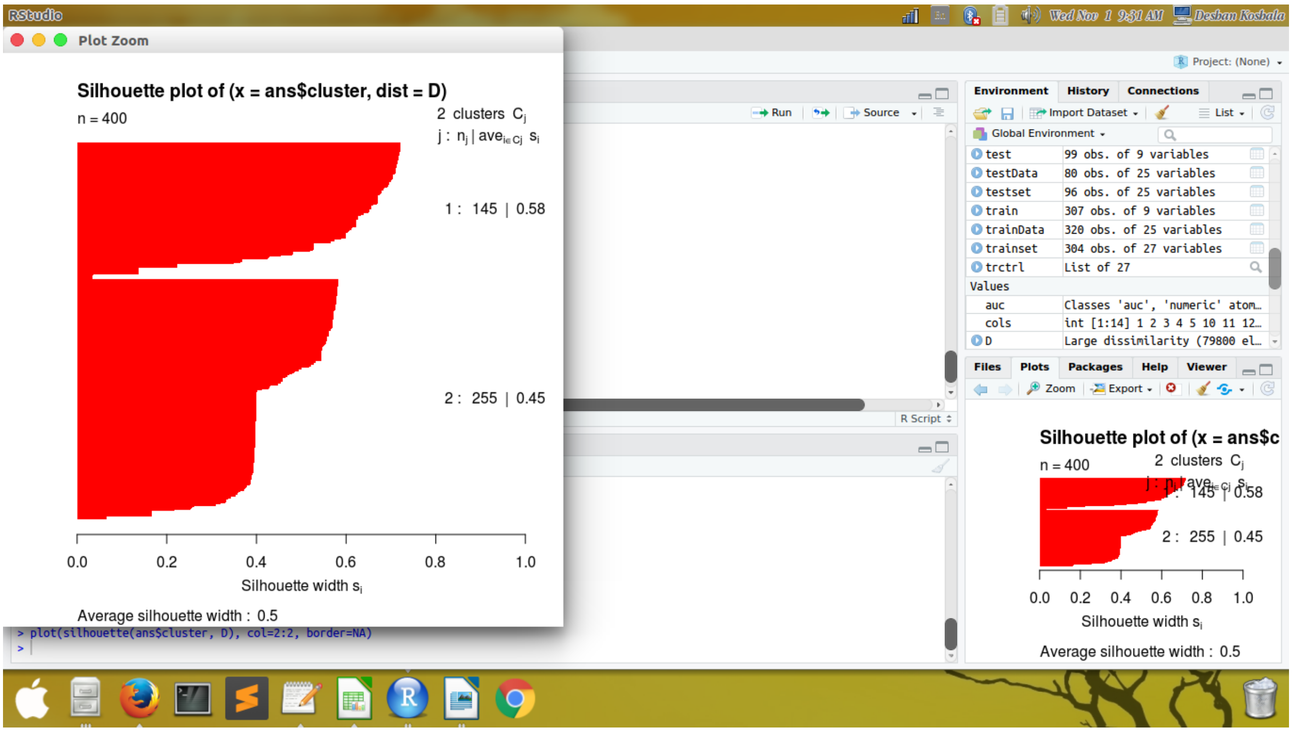
```
> plot(silhouette(ans$cluster, D), col=1:1, border=NA)
```



```
test        99 obs. of 9 variables
testData    80 obs. of 25 variables
testset     96 obs. of 25 variables
train       307 obs. of 9 variables
trainData   320 obs. of 25 variables
trainset    304 obs. of 27 variables
trctrl      List of 27
Values
  auc       Classes 'auc', 'numeric' atom…
  cols      int [1:14] 1 2 3 4 5 10 11 12…
D           Large dissimilarity (79800 el…
```

```
> plot(silhouette(ans$cluster, D), col=1:1, border=NA)
>
```

```
> plot(silhouette(ans$cluster, D), col=2:2, border=NA)
```



test	99 obs. of 9 variables
testData	80 obs. of 25 variables
testset	96 obs. of 25 variables
train	307 obs. of 9 variables
trainData	320 obs. of 25 variables
trainset	304 obs. of 27 variables
trctrl	List of 27
Values	
auc	Classes 'auc', 'numeric' atom…
cols	int [1:14] 1 2 3 4 5 10 11 12…
D	Large dissimilarity (79800 el…

```
> plot(silhouette(ans$cluster, D), col=2:2, border=NA)
>
```

B) K means clustering with k=actual number of classes-1

```
cols <- c(1:5, 10:18)
ans <- kmeans(kidney_noc[, cols], 1)
ans
```

```
K-means clustering with 1 clusters of sizes 400

Cluster means:
           age          bp        sg  al    su         bgr          bu           sc         sod
pot         hemo        pcv
1 51.48337596 76.46907216 1.0177125 0.9 0.395 148.0365169 57.4257218 3.072454295 137.528754
4.62724368 12.52643681 38.8844984
         wbcc        rbcc
1 8406.122449 4.707434963

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [67] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[133] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[199] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[265] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[331] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[397] 1 1 1 1

Within cluster sum of squares by cluster:
[1] 2543757430
 (between_SS / total_SS =   0.0 %)
```

```
Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
"size"         "iter"
[9] "ifault"
```

C) K means clustering with k=actual number of classes+1

```
> cols <- c(1:5, 10:18)
> ans <- kmeans(kidney_noc[, cols], 3)
> ans
K-means clustering with 3 clusters of sizes 12, 111, 277

Cluster means:
           age          bp          sg            al            su          bgr           bu
sc          sod         pot         hemo
1 52.25000000 71.66666667 1.012500000 2.4166666667 0.5833333333 169.6697098 98.33333333
4.758333333 132.9214590 4.137874000 10.00000000
2 53.21591668 77.80068728 1.017162162 0.9459459459 0.6126126126 155.9489068 53.93742641
2.507565910 138.6257663 4.359782919 12.71987160
3 50.75589759 76.14351110 1.018158845 0.8158844765 0.2996389892 143.9286700 57.05138768
3.225783040 137.2887510 4.755621227 12.55837175
           pcv         wbcc         rbcc
1 29.41666667 17433.333333 3.667286250
2 38.84164405 10299.099099 4.659228378
3 39.31182985  7256.494511 4.771812996

Clustering vector:
  [1] 3 3 3 3 3 3 3 3 2 2 3 3 2 3 2 3 3 3 2 3 2 3 3 3 3 3 2 2 3 3 3 3 2 3 3 2 3 3 3 2 2 3 2
3 3 3 3 3 1 1 2 3 3 3 2 3 2 3 3 3 3 3 3 3 3 3
 [67] 3 3 3 3 3 1 3 3 3 3 3 3 3 2 2 3 3 3 3 3 3 3 2 3 2 2 3 3 2 3 3 2 3 2 3 2 3 2 3 3 3 2 2
3 2 2 3 2 2 3 3 2 3 3 3 3 3 1 3 3 3 1 3 3 1 2
[133] 2 3 2 3 3 2 3 3 3 3 3 3 2 3 3 1 3 3 3 3 3 1 3 3 3 3 3 2 2 2 3 3 2 3 3 3 3 3 2 2 2 2 3
3 2 3 2 3 3 1 3 3 3 3 3 2 3 3 1 2 3 3 3 3 2 3
[199] 1 2 3 3 3 3 3 3 3 2 3 3 2 3 2 3 3 3 3 2 3 3 2 3 3 2 3 2 3 3 3 1 2 3 3 3 3 3 3 2 3 3 3
3 2 2 2 3 3 3 2 3 2 2 2 3 3 2 2 2 3 3 2 3 3 3
[265] 3 2 3 2 3 3 3 2 3 3 3 3 3 3 2 3 3 2 3 2 3 3 2 3 2 3 3 3 2 3 3 2 3 3 3 2 2 3 3 3 3 2 3
3 3 3 2 3 3 3 2 3 3 3 3 3 2 2 3 3 3 3 3 2 3 3
[331] 3 3 2 3 3 2 2 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 3 3 3 3 3 2 3 3 3 3 2 3 2 2 2 3 2 2 2 3 2
2 3 2 2 3 3 3 3 3 3 3 2 3 3 3 2 3 3 3 3 3 2 3
[397] 3 3 3 3

Within cluster sum of squares by cluster:
[1] 140094535.8 138132948.2 523753530.0
 (between_SS / total_SS =  68.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
"size"         "iter"
[9] "ifault"
```

D) Density based clustering with the radius parameter

** install.packages(“dbscan”)

```
 library (cluster)
 library (dbscan)
 cols <- c(1:5, 10:18)
 ans <- clara (x=kidney_noc[, cols], k=2)
 ans
Call:    clara(x = kidney_noc[, cols], k = 2)
```

```
Medoids:
     age bp    sg al su bgr        bu        sc         sod       pot hemo        pcv
wbcc     rbcc
[1,]  55 80 1.010  0  0 146 57.425722 3.072454 137.528754 4.627244  9.8 38.884498
8406.122449 4.707435
[2,]  52 80 1.025  0  0  99 25.000000 0.800000 135.000000 3.700000 15.0 52.000000
6300.000000 5.300000
Objective function:      1136.183539
Clustering vector:       int [1:400] 1 2 1 2 2 1 1 2 1 1 1 2 1 1 1 2 1 1 ...
Cluster sizes:           279 121
Best sample:
 [1]   2  10  38  65  74  76  77  83  90  97 105 123 133 137 144 153 156 160 169 173 175 178
186 203 204 212 219 229 245 257 294 299 304
[34] 311 314 321 346 351 353 354 379 382 391 397

Available components:
 [1] "sample"     "medoids"    "i.med"      "clustering" "objective"  "clusinfo"   "diss"
"call"       "silinfo"    "data"
```

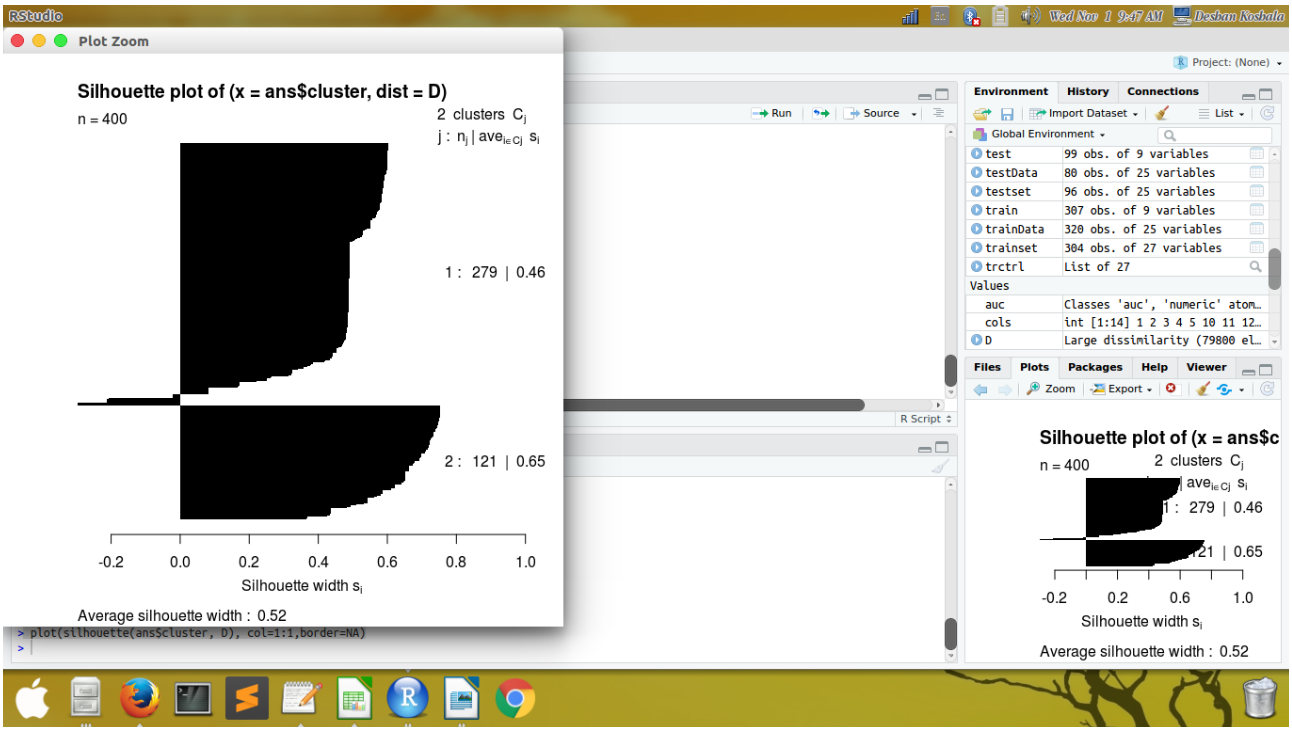
```
ans[["clusinfo"]]
```

```
     size    max_diss      av_diss   isolation
[1,]  279 17994.21259 1251.2729450 8.540413057
[2,]  121  4100.26498  870.8120988 1.946067737
```

```
dbscan(kidney_noc[, cols], eps=11047, minPts = 121 )
```

```
DBSCAN clustering for 400 objects.
Parameters: eps = 11047, minPts = 121
The clustering contains 1 cluster(s) and 0 noise points.

  1
400

Available fields: cluster, eps, minPts
```

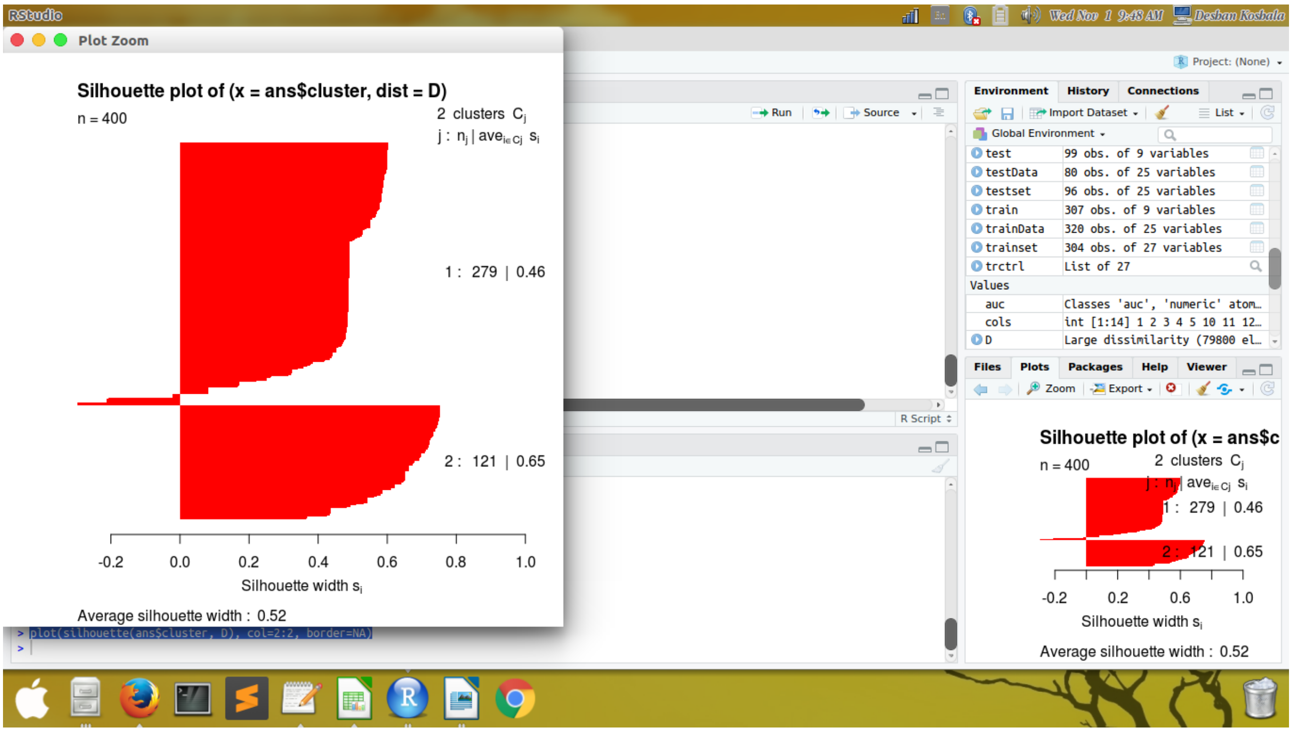
```
 library(cluster)
 D<- daisy(kidney_noc[, cols])

 plot(silhouette(ans$cluster, D), col=1:1,border=NA)
```



```
> plot(silhouette(ans$cluster, D), col=1:1,border=NA)
> 
```

test	99 obs. of 9 variables
testData	80 obs. of 25 variables
testset	96 obs. of 25 variables
train	307 obs. of 9 variables
trainData	320 obs. of 25 variables
trainset	304 obs. of 27 variables
trctrl	List of 27
Values	
auc	Classes 'auc', 'numeric' atom…
cols	int [1:14] 1 2 3 4 5 10 11 12…
D	Large dissimilarity (79800 el…

```
plot(silhouette(ans$cluster, D), col=2:2, border=NA)
```



test	99 obs. of 9 variables
testData	80 obs. of 25 variables
testset	96 obs. of 25 variables
train	307 obs. of 9 variables
trainData	320 obs. of 25 variables
trainset	304 obs. of 27 variables
trctrl	List of 27
Values	
auc	Classes 'auc', 'numeric' atom…
cols	int [1:14] 1 2 3 4 5 10 11 12…
D	Large dissimilarity (79800 el…

```
> plot(silhouette(ans$cluster, D), col=2:2, border=NA)
>
```

The cluster with highest (between_SS / total_SS) would be the best cluser, accordingly out of the three cases A, B and C , C is the better cluster.