2,

Because the variable “gamma” has the miss value from observation 66 to 105, I treat the miss value firstly by choosing the median value of the 10 nearest observation’s “gamma” values.

distance<-as.matrix(dist(primate))

for(n in c(10,10)){

for(i in c(66:105)){

primate[i,"gamma"]<-apply(data.frame(primate[c(as.integer(names(sort(distance[i,])[2:n]))),"gamma"]),2,median,na.rm=T)

}

}

a,

Single-linkage:

I try to divide the data to different groups from 2 to 10 and I got the different misclassification rate.

0.7509158(2) 0.2573260 (3) 0.1639194(4) 0.1686813(5) 0.1708791(6) 0.1752747(7) 0.1706960 (8) 0.1756410(9) 0.1782051(10)

The group (4) has the minimum misclassification rate 0.1639194. The misclassification rate of the group (5) is 0.1686813.

The group (4)

1 2 3 4

Gorilla 0 0 0 14

Homo 0 0 40 0

Hylobates 15 1 0 0

Pan 0 0 0 20

Pongo 0 0 15 0

The group (5)

1 2 3 4 5

Gorilla 0 0 0 0 14

Homo 0 0 0 40 0

Hylobates 13 2 1 0 0

Pan 0 0 0 0 20

Pongo 0 0 0 15 0

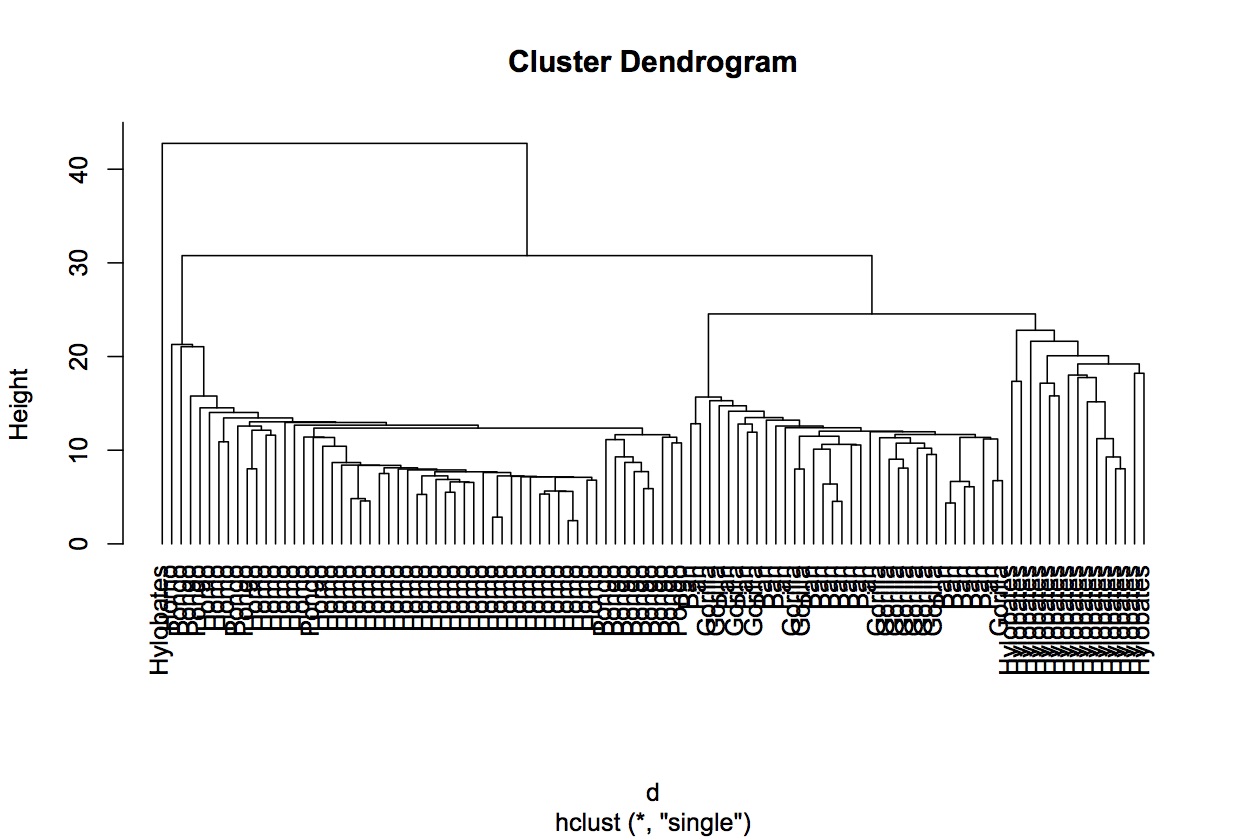
From the above two tables, we also can find that the group(4) is better.

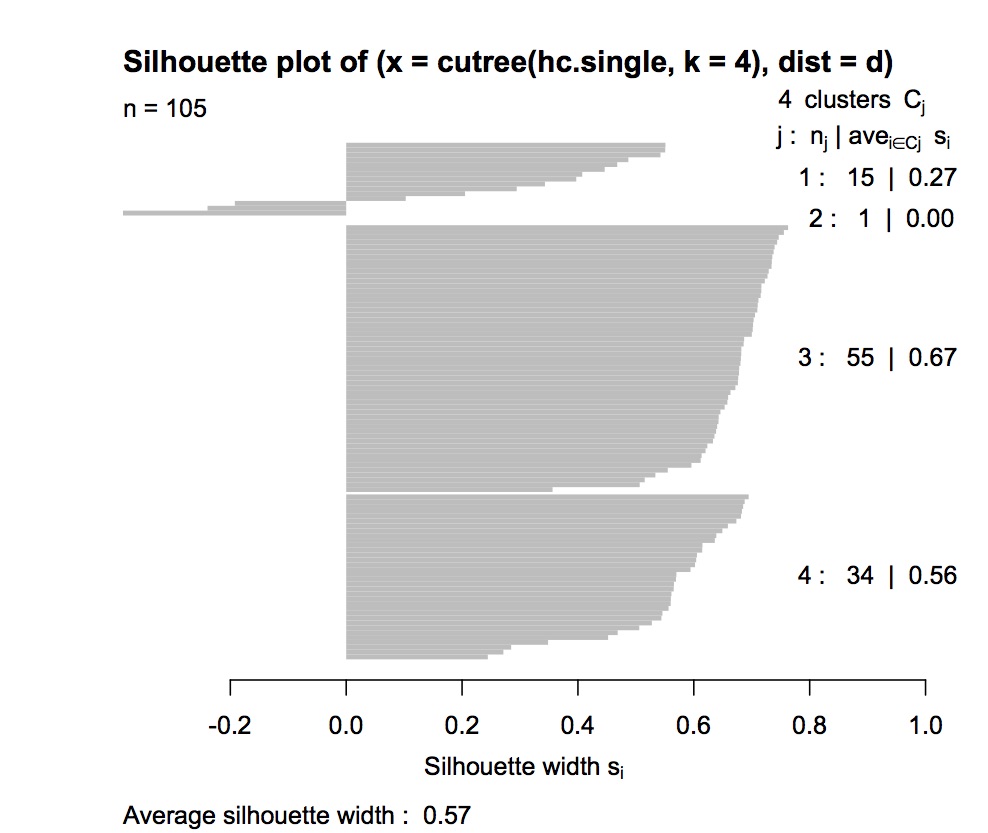
I also plot the silhouette plots of the different groups and I get the average widths of the silhouette plots.

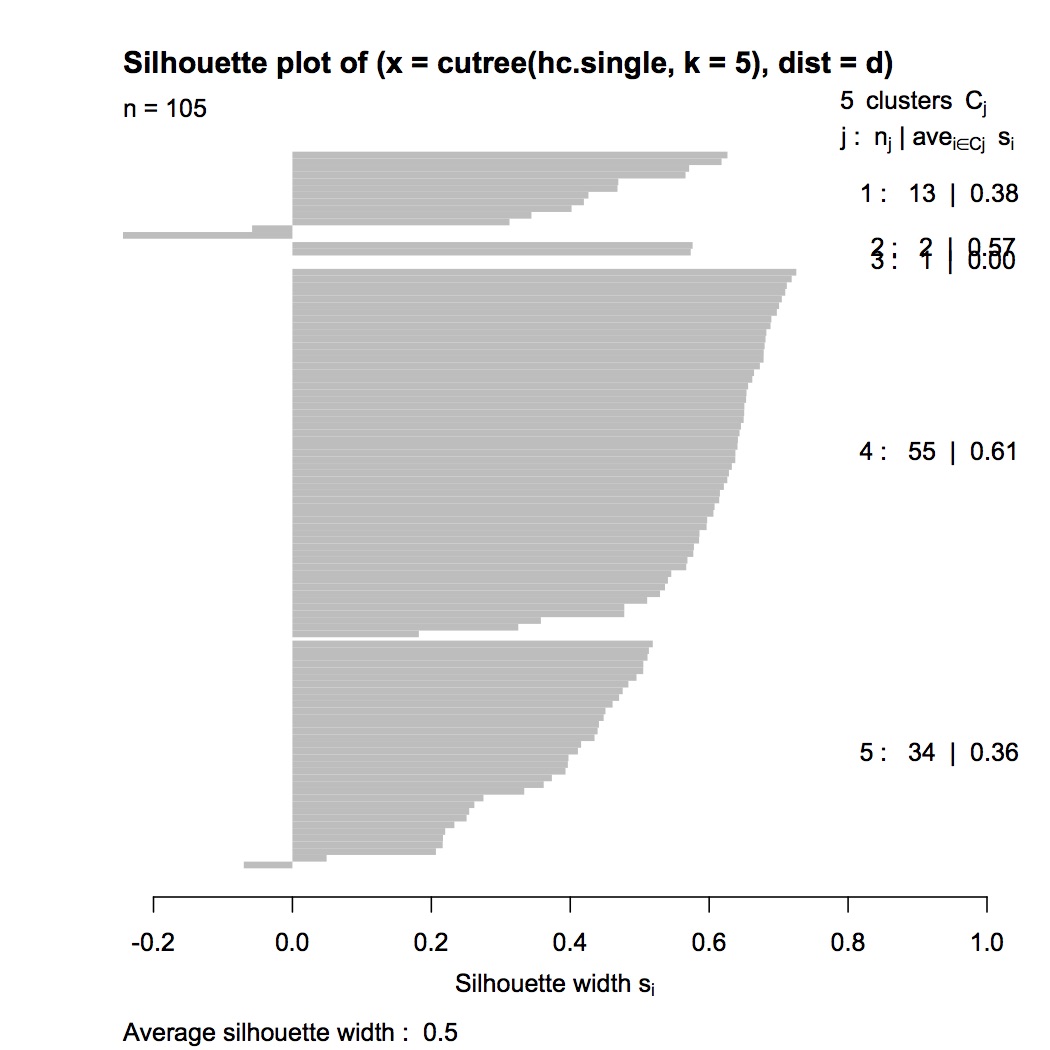
0.5588978(2) 0.5553507(3) 0.5694269(4) 0.4954185(5) 0.4729187(6) 0.3306103(7) 0.2732398(8) 0.2817304(9) 0.2694172(10)

The group(4) also has the maximum average widths of the silhouette plots.

For the single-linkage I divide the 4 groups.







Average-linkage:

I try to divide the data to different groups from 2 to 10 and I got the different misclassification rate.

0.75091575(2) 0.25732601(3) 0.18113553(4) 0.18772894(5) 0.17527473(6) 0.17967033(7) 0.09157509(8) 0.09304029(9) 0.09450549(10)

The group (8) has the minimum misclassification rate 0.09157509. The misclassification rate of the group (4) is 0.18113553 and the group (5) is 0.18772894. The group(4) is better than group (5).

The group(4)

1 2 3 4

Gorilla 0 14 0 0

Homo 0 0 0 40

Hylobates 13 2 1 0

Pan 0 20 0 0

Pongo 0 0 0 15

The group(5)

1 2 3 4 5

Gorilla 0 14 0 0 0

Homo 0 0 0 0 40

Hylobates 4 2 1 9 0

Pan 0 20 0 0 0

Pongo 0 0 0 0 15

The group(8)

1 2 3 4 5 6 7 8

Gorilla 0 0 0 0 0 0 14 0

Homo 0 0 0 0 1 38 0 1

Hylobates 4 2 1 9 0 0 0 0

Pan 0 0 0 0 0 0 20 0

Pongo 0 0 0 0 14 1 0 0

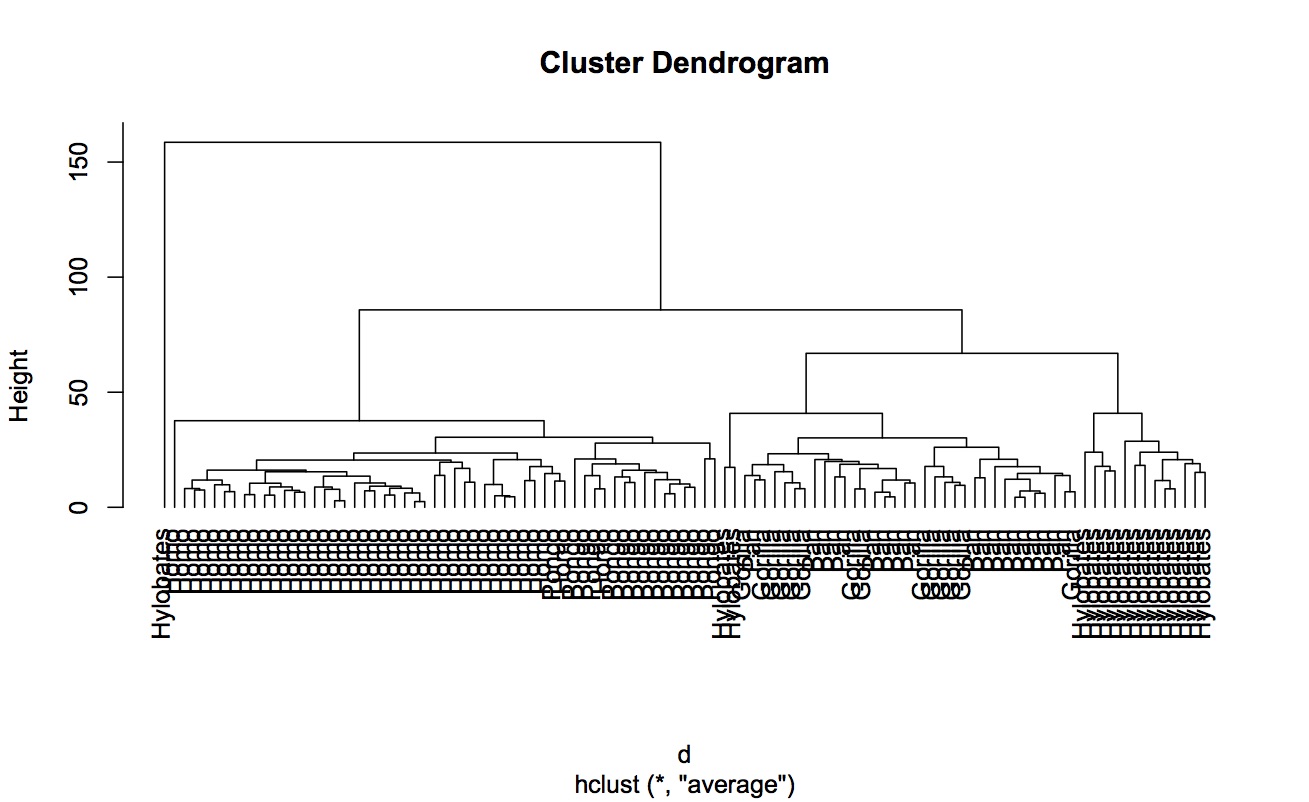
From the above three tables, we can find that the group(4) is still better.

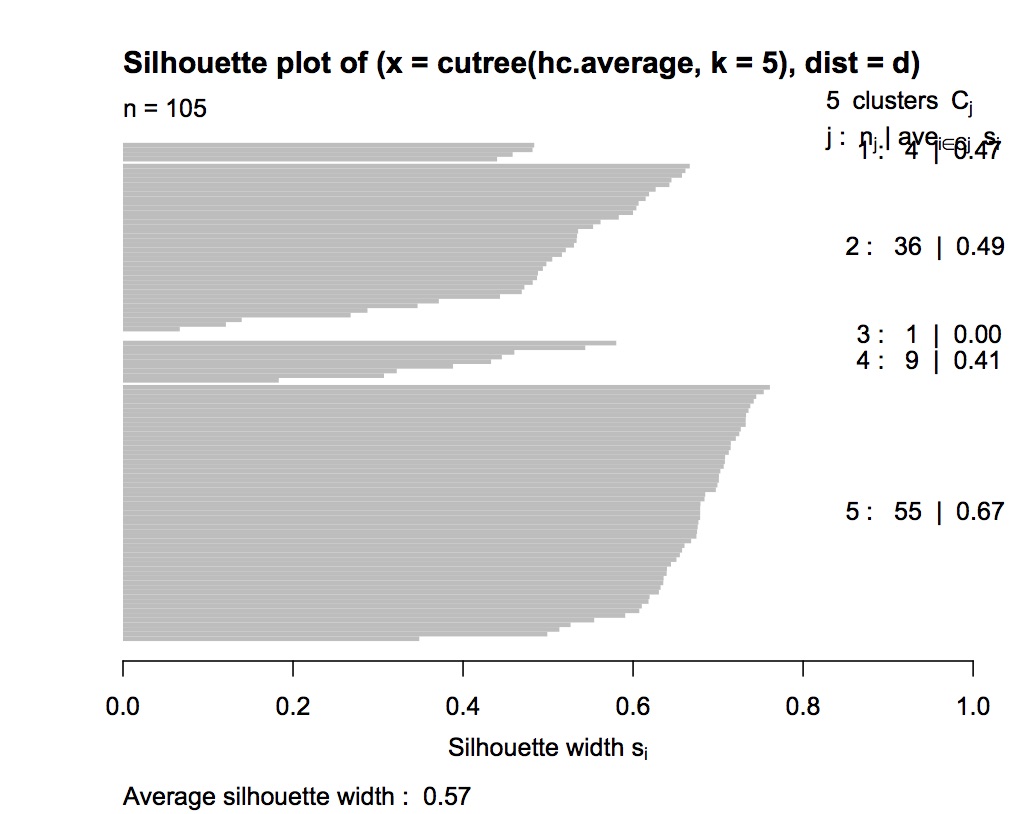
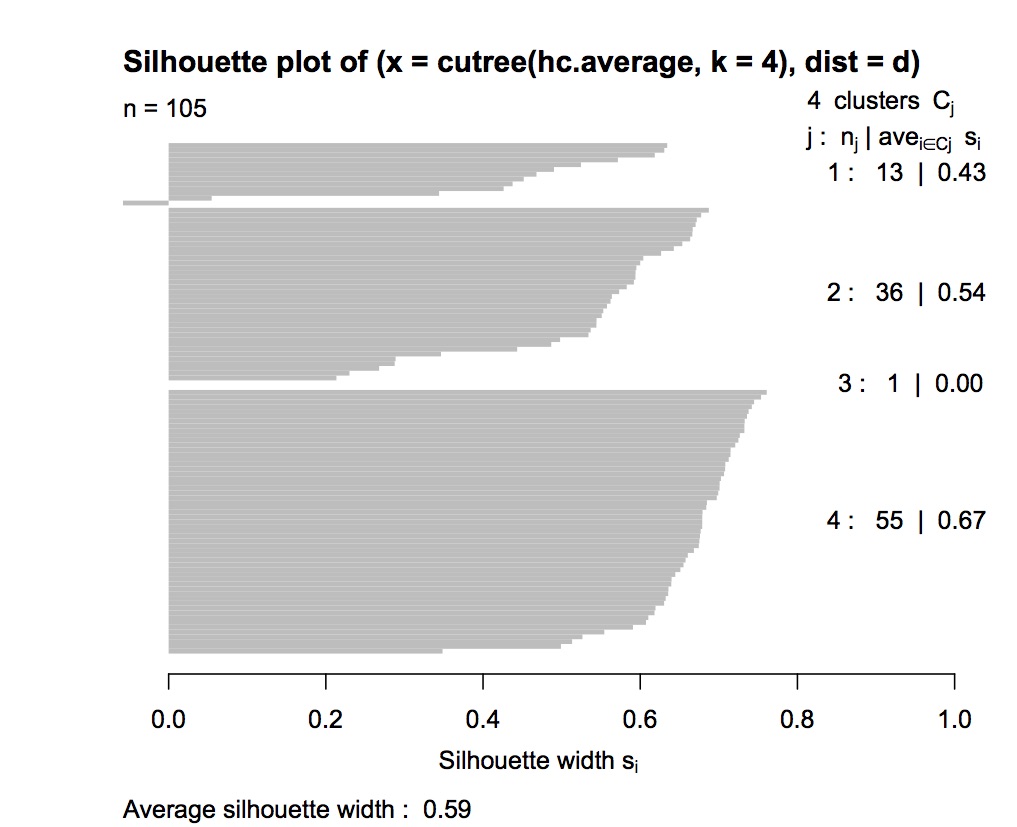
I also plot the silhouette plots of the different groups and I get the average widths of the silhouette plots.

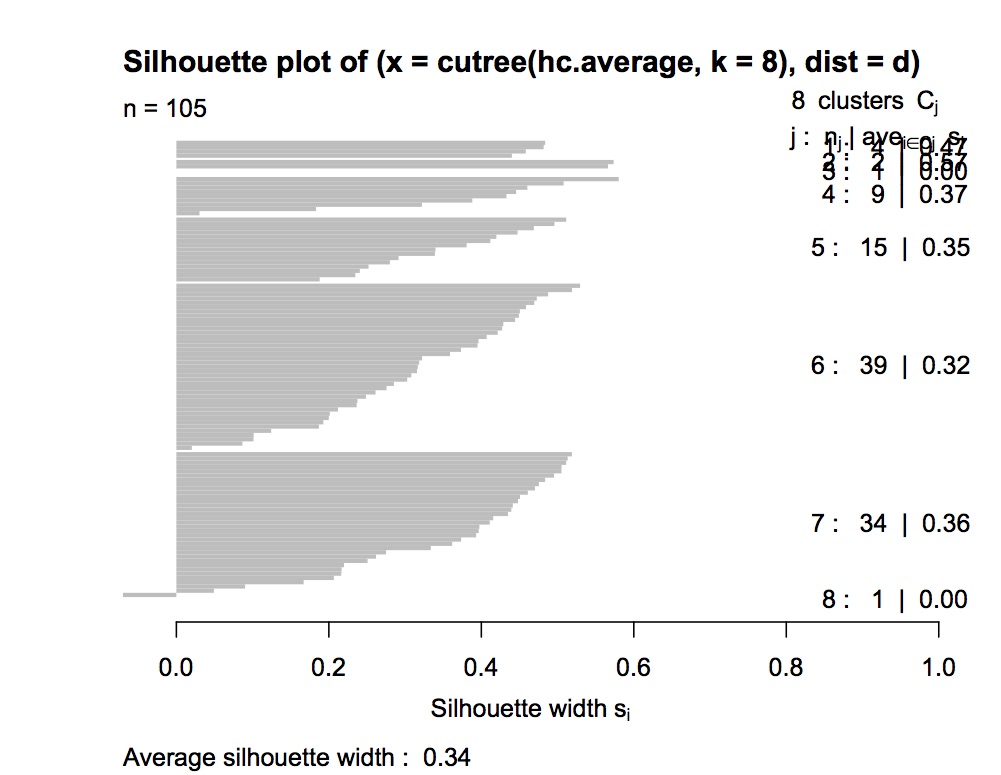
0.5588978(2) 0.5553507(3) 0.5867011(4) 0.5706454(5) 0.4960227(6) 0.3537143(7) 0.3440094(8) 0.3249936(9) 0.3022507(10)

The group(4) also has the maximum average widths of the silhouette plots.

In summary for the average-linkage, I still choose 4 groups.







Complete-linkage:

I try to divide the data to different groups from 2 to 10 and I got the different misclassification rate.

0.6930403(2) 0.2397436(3) 0.1869963(4) 0.1752747(5) 0.1760073(6) 0.1419414(7) 0.1382784(8) 0.1282051(9) 0.1265568(10)

The group (10) has the minimum misclassification rate 0.1265568. The misclassification rate of the group (4) is 0.1869963 and the group (5) is 0.1752747.

The group(4)

1 2 3 4

Gorilla 0 14 0 0

Homo 0 0 0 40

Hylobates 5 2 9 0

Pan 0 20 0 0

Pongo 0 0 0 15

The group(5)

1 2 3 4 5

Gorilla 0 6 0 0 8

Homo 0 0 0 40 0

Hylobates 5 2 9 0 0

Pan 0 4 0 0 16

Pongo 0 0 0 15 0

The group(10)

1 2 3 4 5 6 7 8 9 10

Gorilla 0 0 0 0 0 0 6 2 6 0

Homo 0 0 0 0 1 8 0 0 0 31

Hylobates 4 2 1 9 0 0 0 0 0 0

Pan 0 0 0 0 0 0 5 11 4 0

Pongo 0 0 0 0 12 3 0 0 0 0

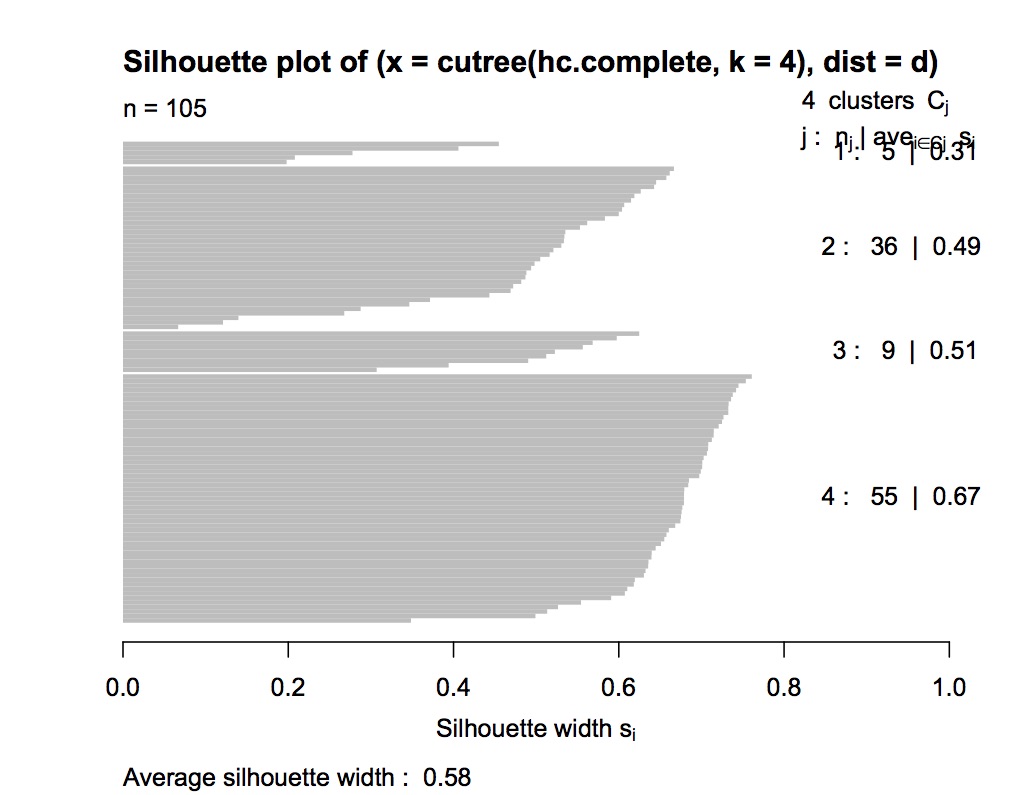
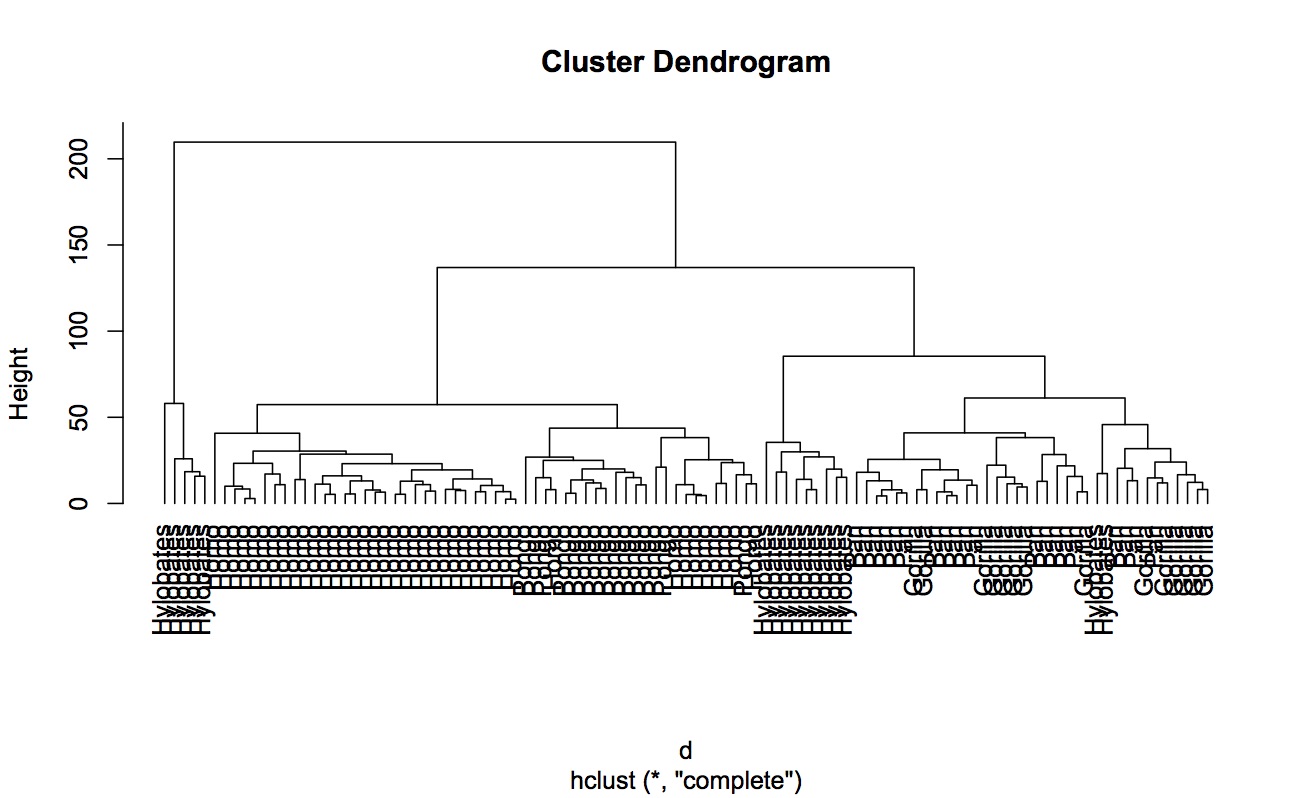
From the above three tables, we can find that the group(4) is still better.

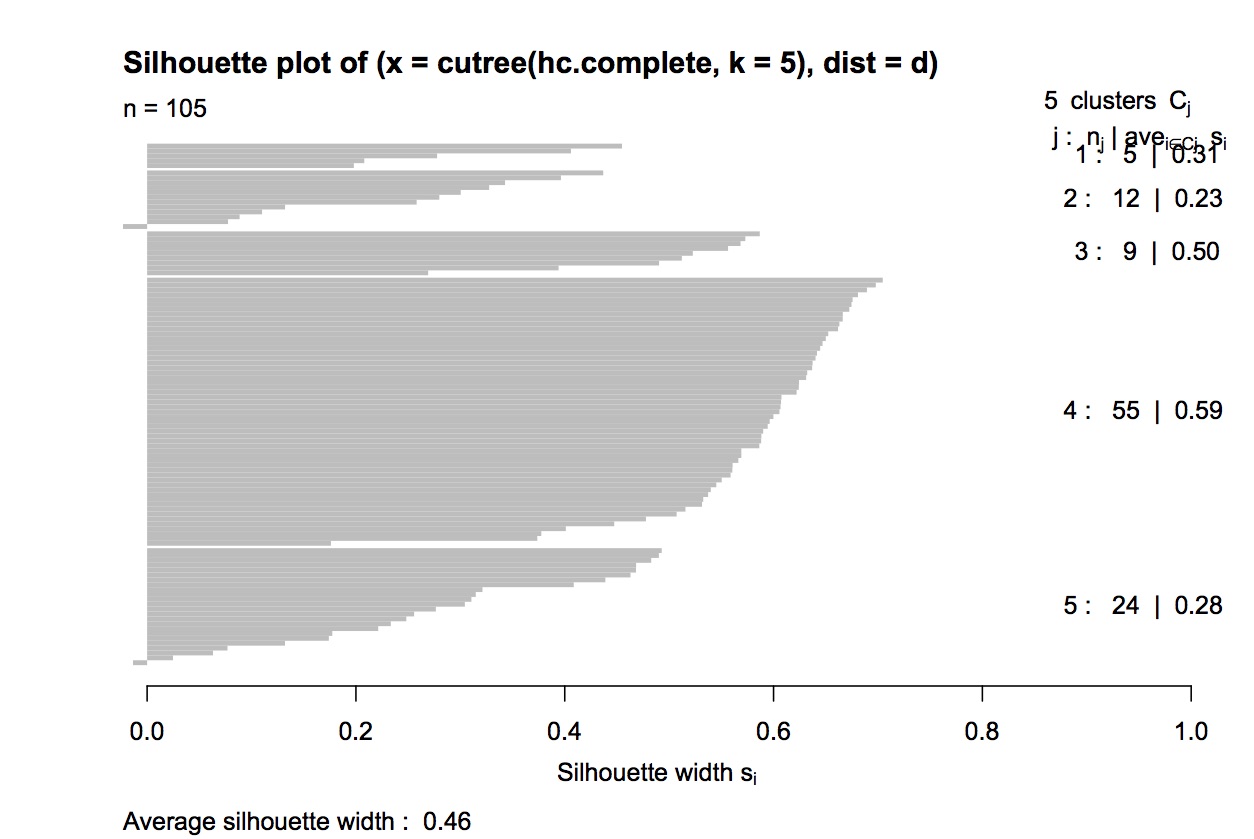
I also plot the silhouette plots of the different groups and I get the average widths of the silhouette plots.

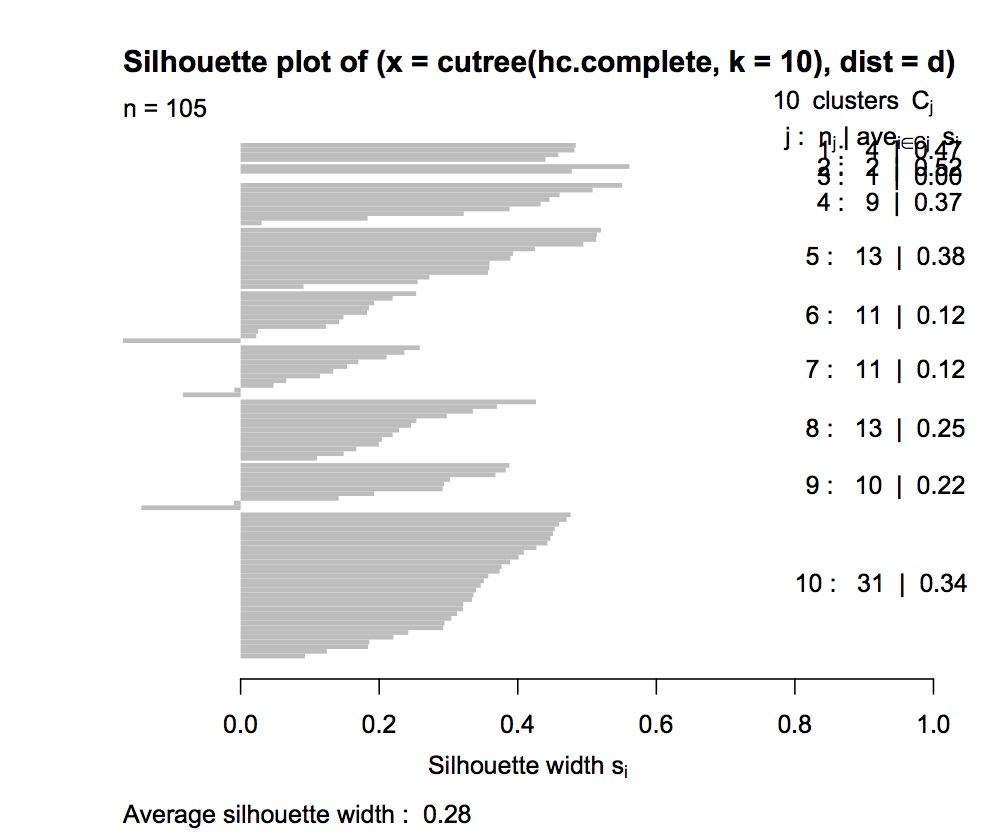
0.4944933(2) 0.5660440(3) 0.5762693(4) 0.4550849(5) 0.4500557(6) 0.3002458(7) 0.2984489 (8) 0.3015802(9) 0.2832376(10)

The group(4) also has the maximum average widths of the silhouette plots.

In summary for the complete-linkage, I still choose 4 groups.







The single-linkage group (4) has the misclassification rate 0.1639194 and the average widths of the silhouette plots is 0.5694269

1 2 3 4

Gorilla 0 0 0 14

Homo 0 0 40 0

Hylobates 15 1 0 0

Pan 0 0 0 20

Pongo 0 0 15 0

The average-linkage group (4) has the misclassification rate 0.18113553 and the average widths of the silhouette plots is 0.5867011

1 2 3 4

Gorilla 0 14 0 0

Homo 0 0 0 40

Hylobates 13 2 1 0

Pan 0 20 0 0

Pongo 0 0 0 15

The complete-linkage group (4) has the misclassification rate 0.1869963 and the average widths of the silhouette plots is 0.5762693

1 2 3 4

Gorilla 0 14 0 0

Homo 0 0 0 40

Hylobates 5 2 9 0

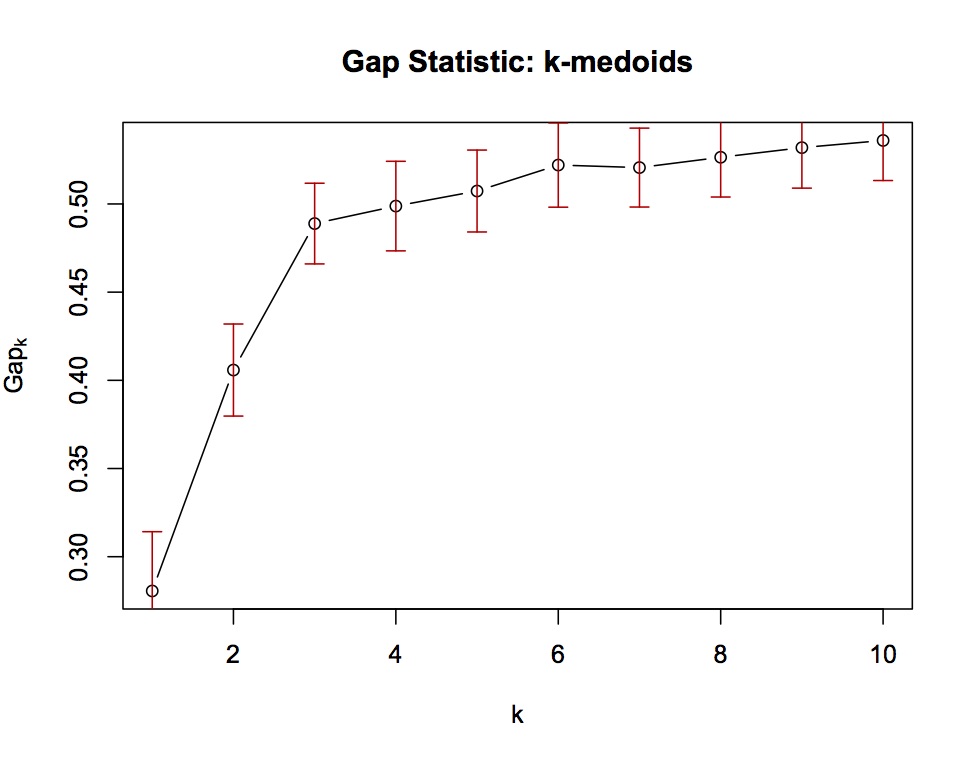
Pan 0 20 0 0

Pongo 0 0 0 15

From all above, the single-linkage is the best and the complete-linkage is worst.

b,

I use K-medoids to treat the data and find the K by clusGap function.



From the plot we can see that k=3 is the elbow point.

The R chooses K=3 automatically.

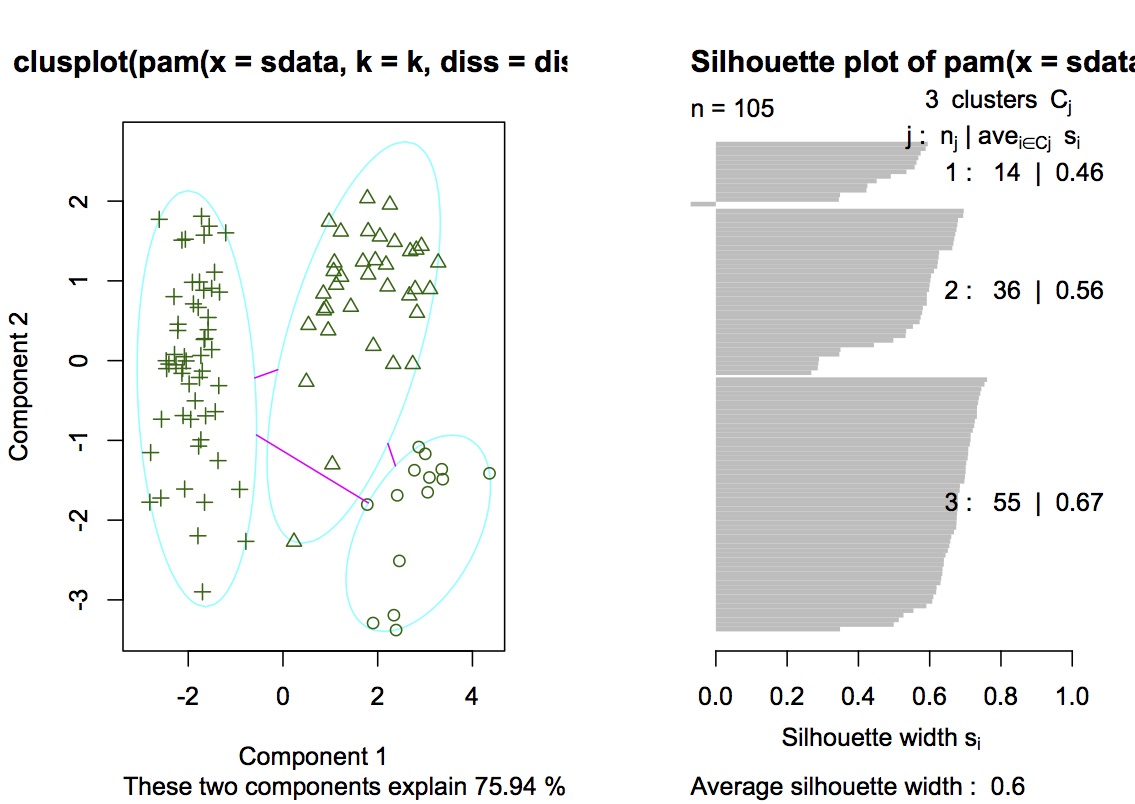
Gorilla Homo Hylobates Pan Pongo

1 0 0 14 0 0

2 14 0 2 20 0

3 0 40 0 0 15

The misclassification rate is 0.1787546.



If K=5

Gorilla Homo Hylobates Pan Pongo

1 0 0 14 0 0

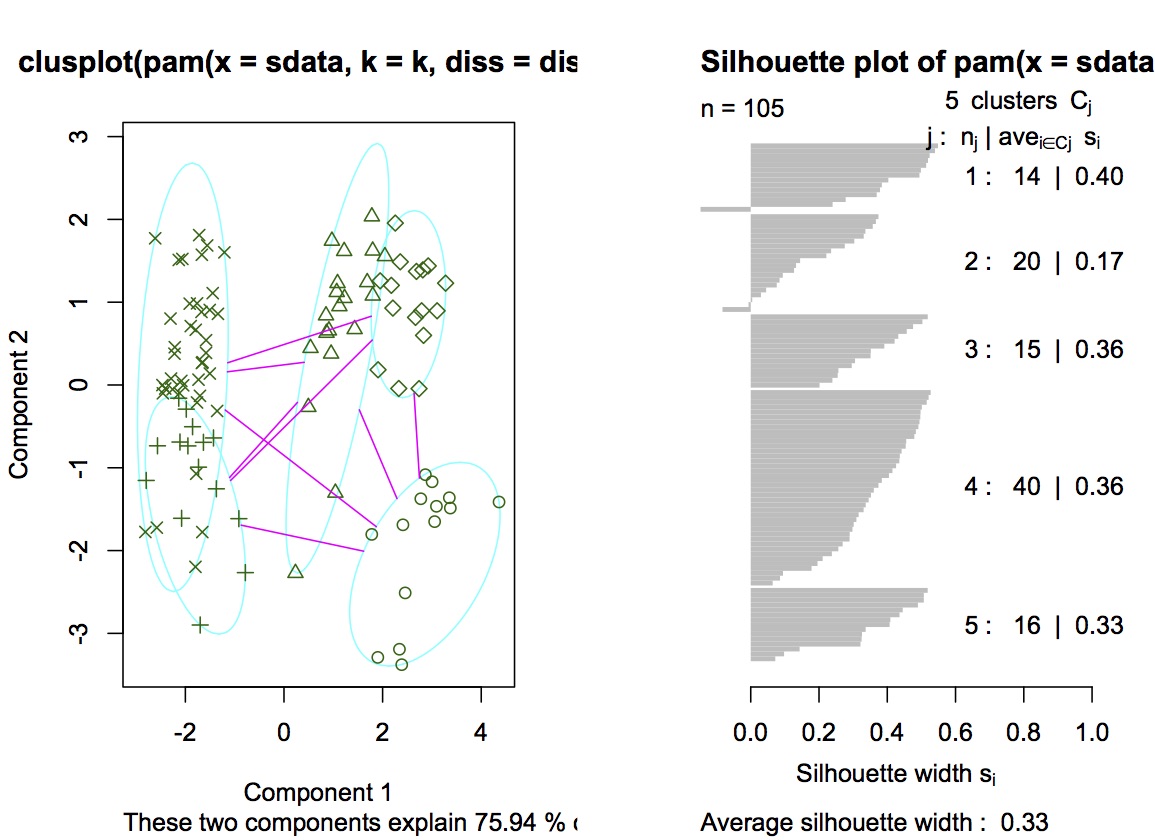
2 12 0 2 6 0

3 0 1 0 0 14

4 0 39 0 0 1

5 2 0 0 14 0

The misclassification rate is 0.06923077.



From all above, I choose K=5.

Comparing with the hierarchical:

K=5

Gorilla Homo Hylobates Pan Pongo

1 0 0 14 0 0

2 12 0 2 6 0

3 0 1 0 0 14

4 0 39 0 0 1

5 2 0 0 14 0

The misclassification rate is 0.06923077.

The single-linkage group (4) has the misclassification rate 0.1639194 and the average widths of the silhouette plots is 0.5694269

1 2 3 4

Gorilla 0 0 0 14

Homo 0 0 40 0

Hylobates 15 1 0 0

Pan 0 0 0 20

Pongo 0 0 15 0

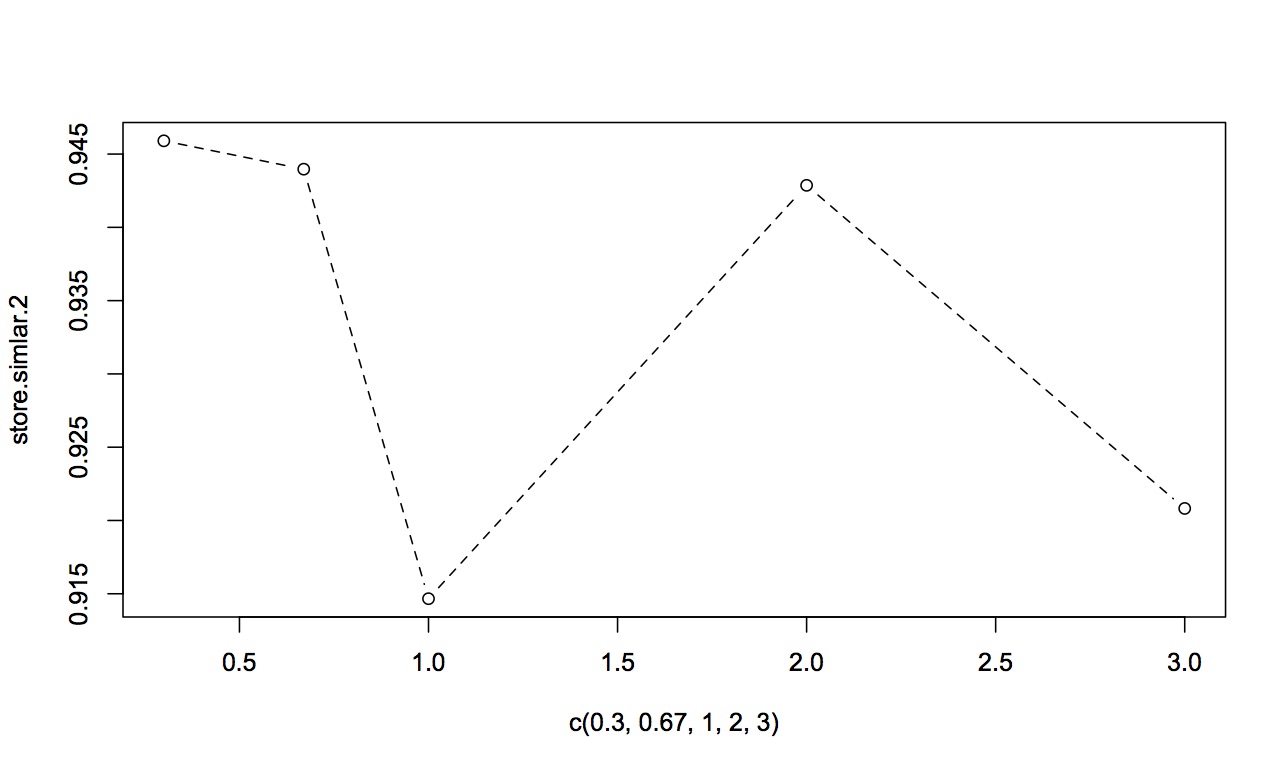
We know that the K-medoids is better than hierarchical.

3,

I assign the radius of som function equal to 0.3, 0.67, 1, 2 and 3.

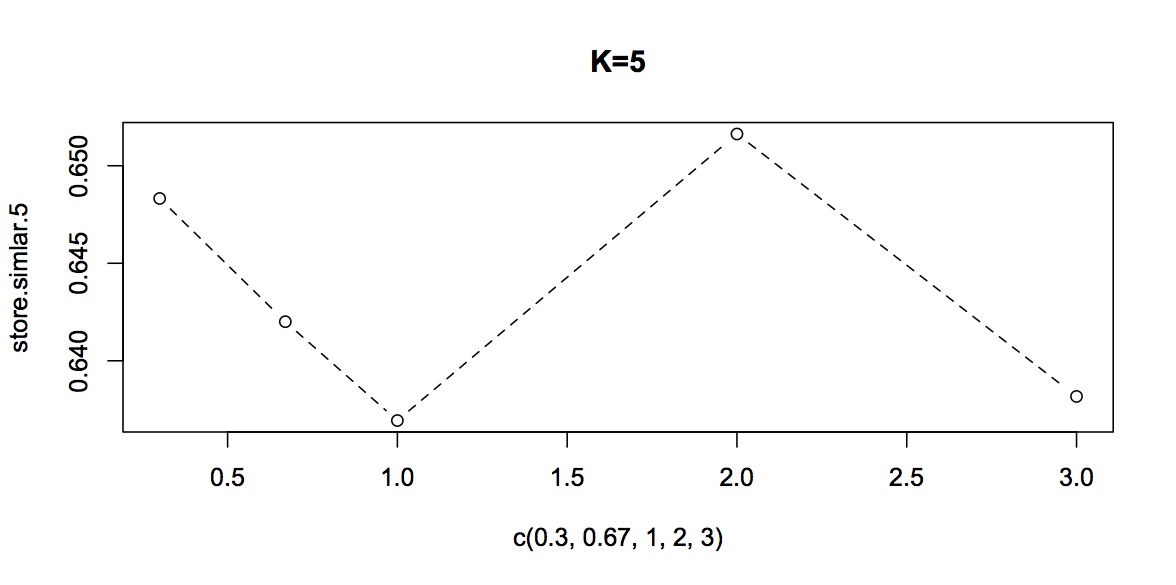
For K=2

The similarity between the K-means and som are 0.9459029, 0.9439690, 0.9146640, 0.9428658 and 0.9208210.

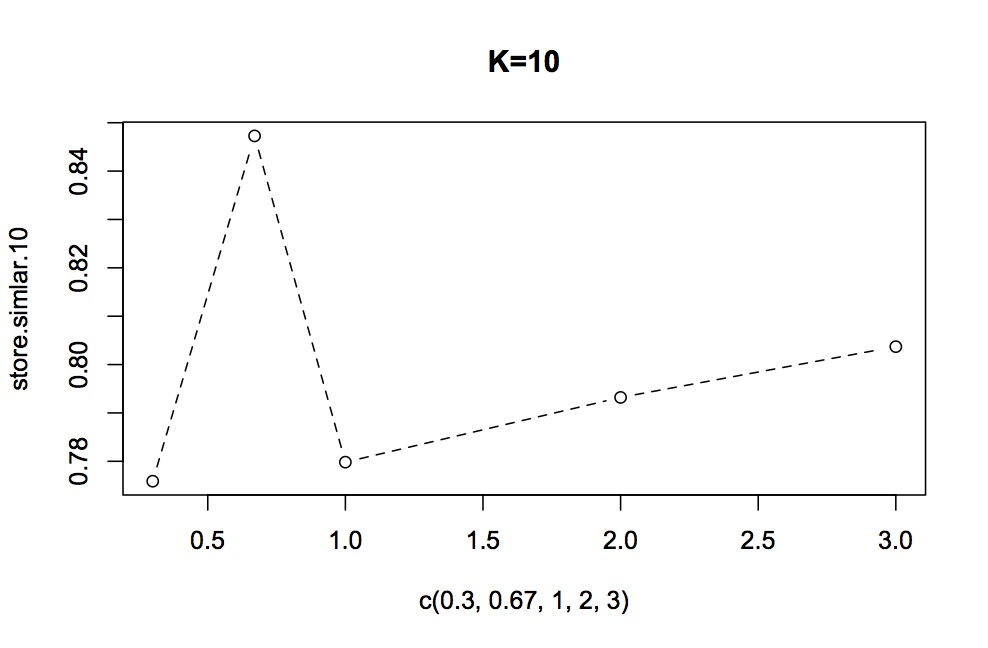


For K=5

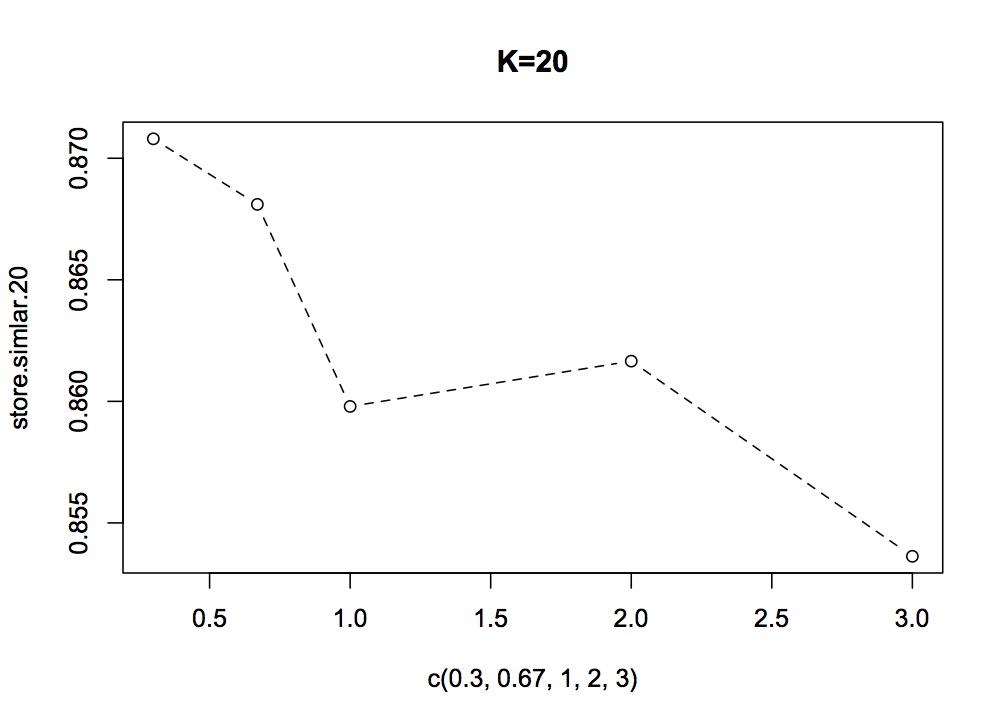
The similarity between the K-means and som are 0.6483224, 0.6420039, 0.6369357, 0.6516280 and 0.6381715.



For K=10

The similarity between the K-means and som are 0.7758879, 0.8472734, 0.7798241, 0.7932066 and 0.8037037.

For K=20

The similarity between the K-means and som are 0.8707991, 0.8681026, 0.8597909, 0.8616545 and 0.8536264.

From the result , we can see that when K equals to20 the assumption is rue but for others the assumption is confirmed.