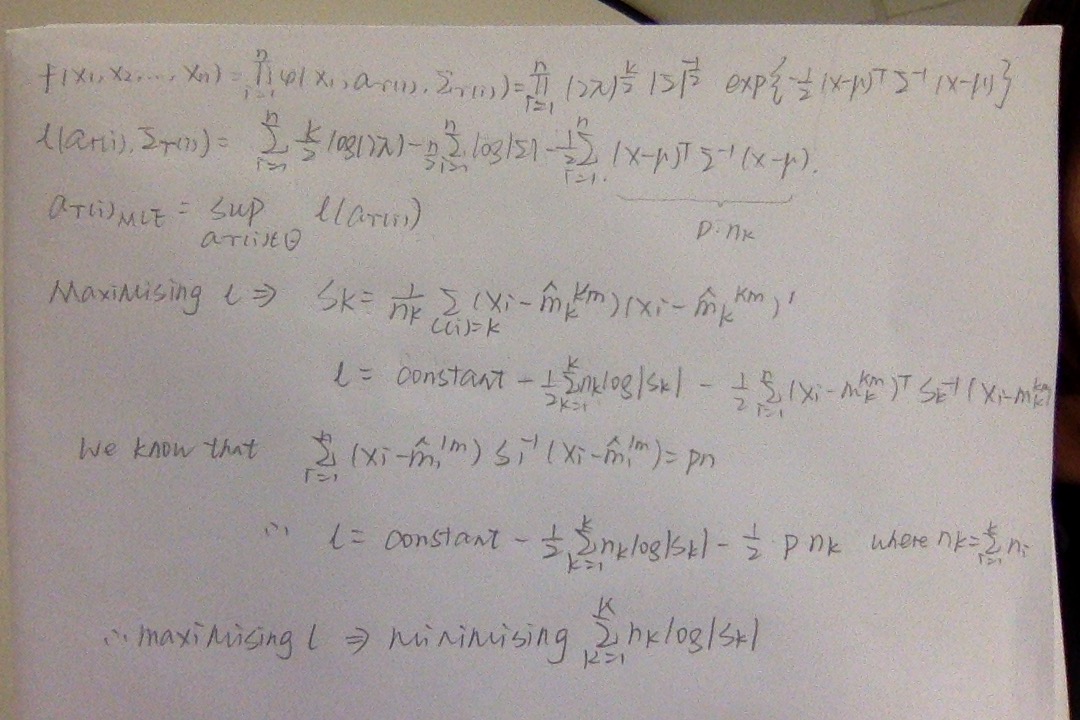
**STAT 3019 Excersice 2**

## **Question 1:**



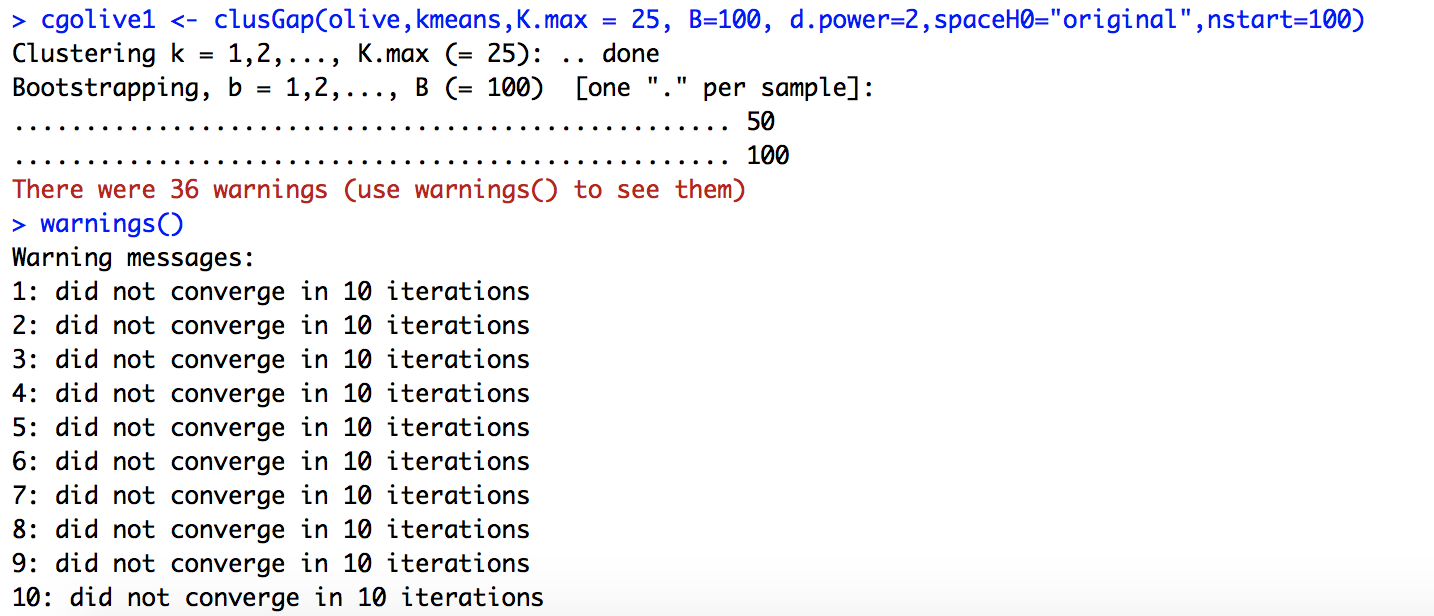
## **Question 2:**

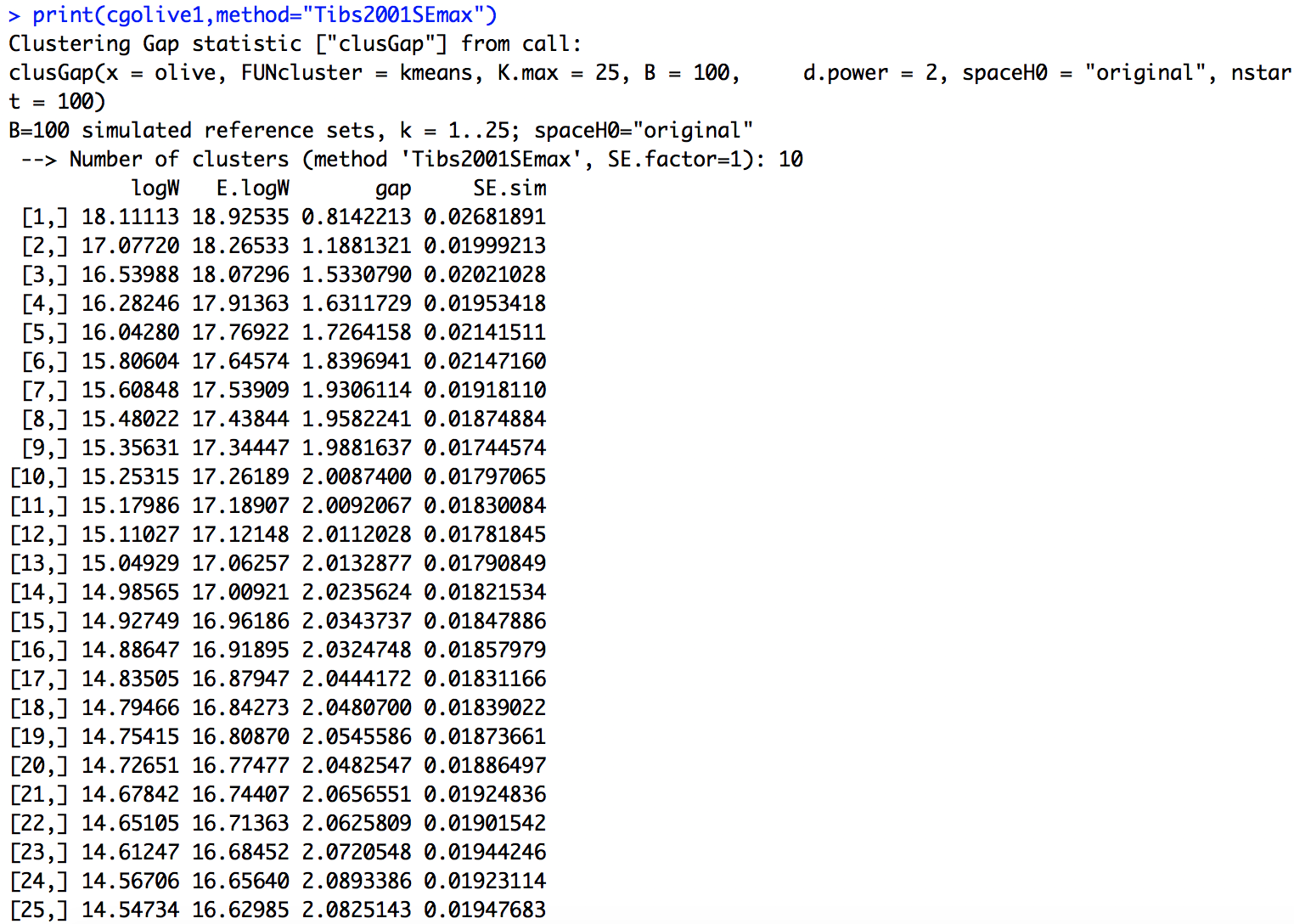
## >R

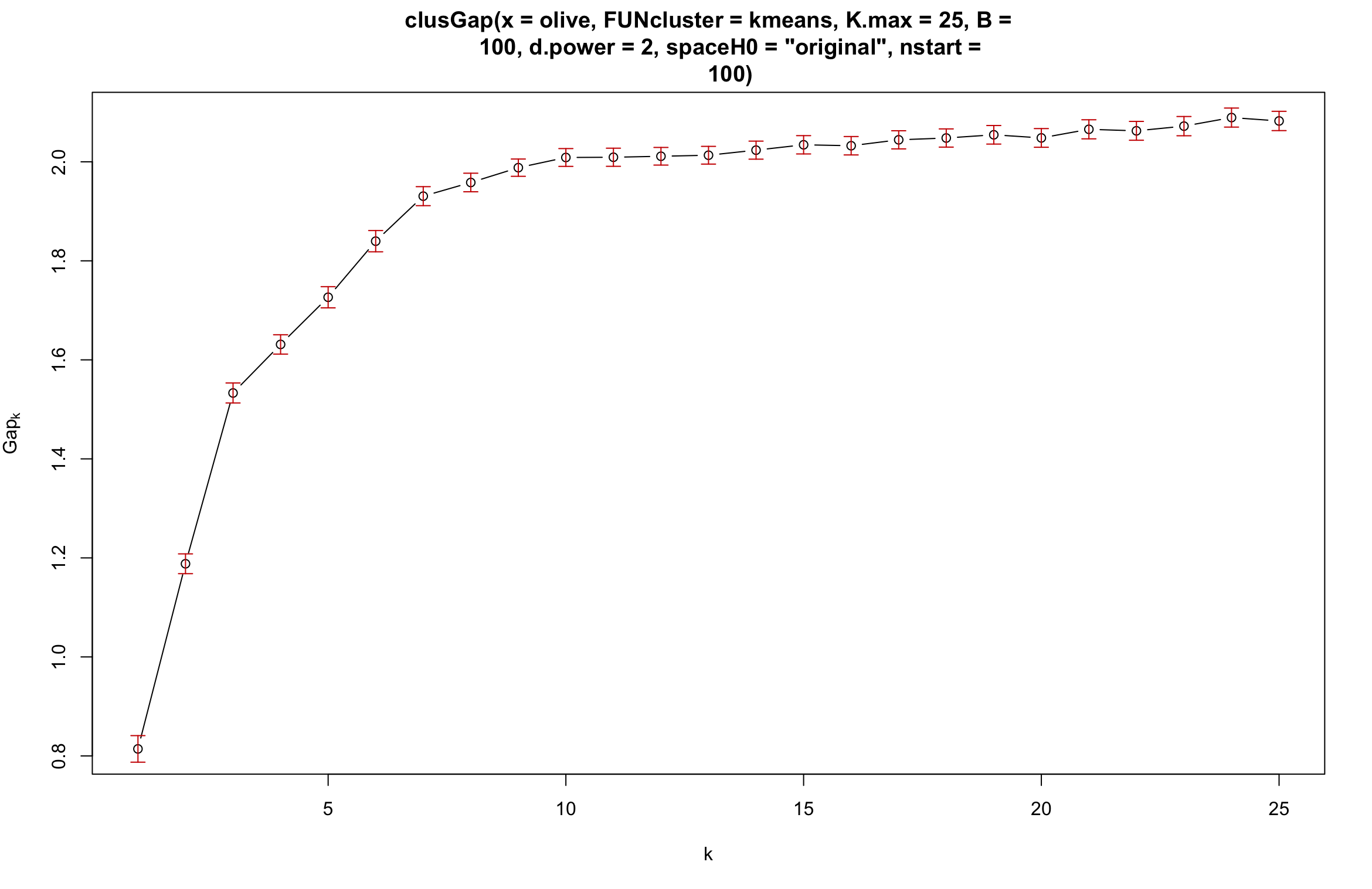
## set.seed(123456)

## cgolive1 <- clusGap(olive,kmeans,K.max = 25, B=100, d.power=2,spaceH0="original",nstart=100)

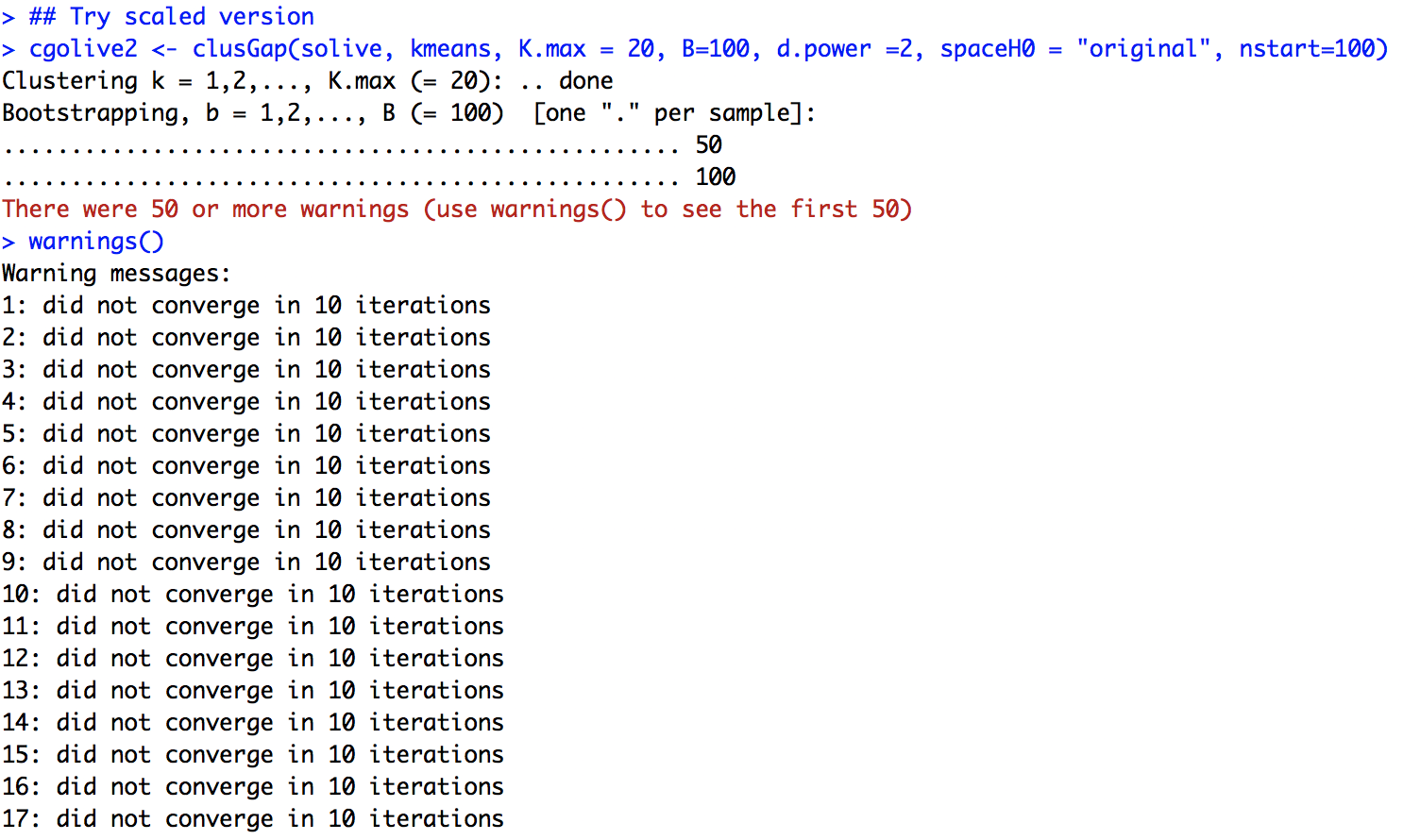
## print(cg1,method="Tibs2001SEmax")

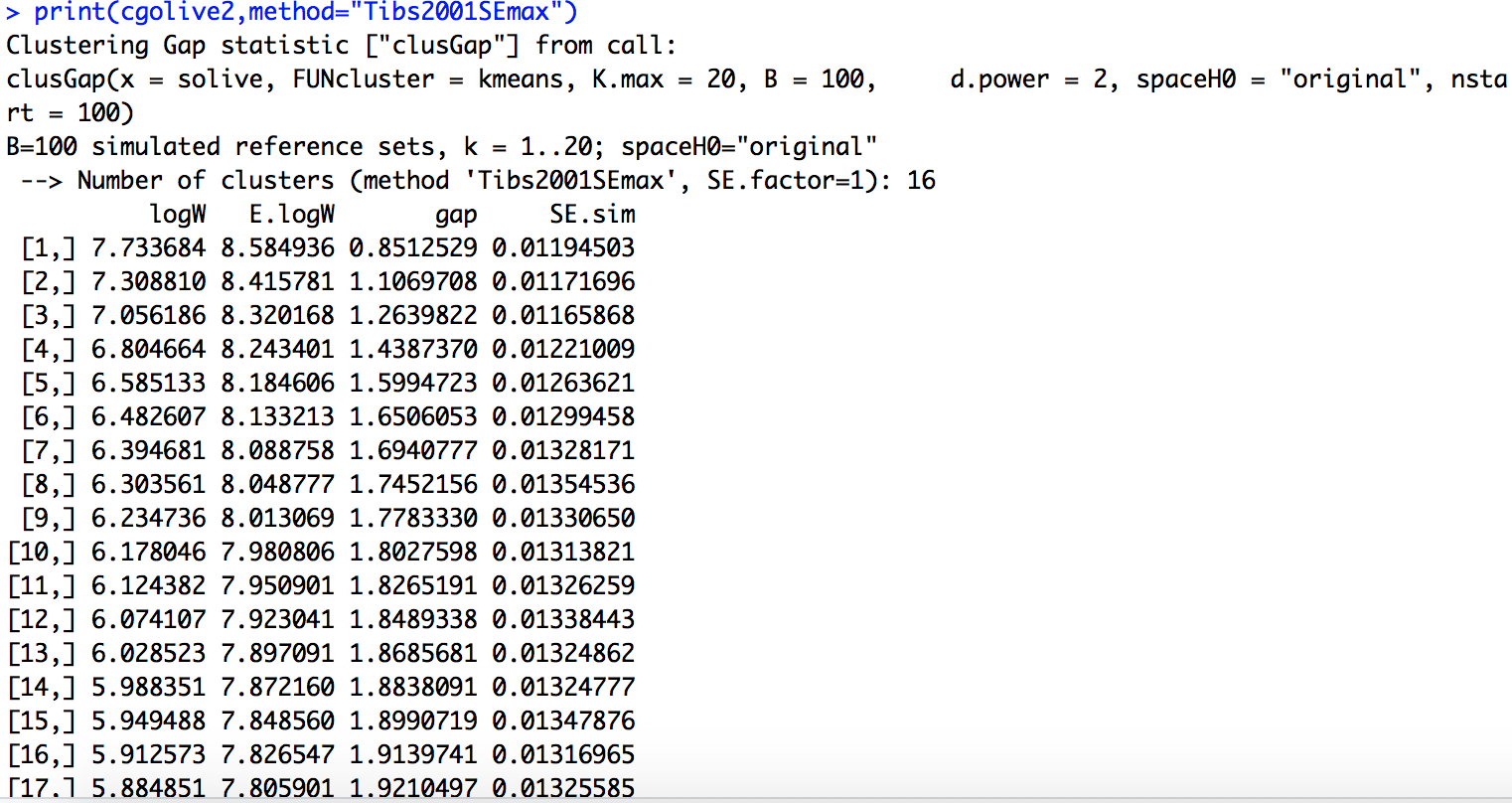


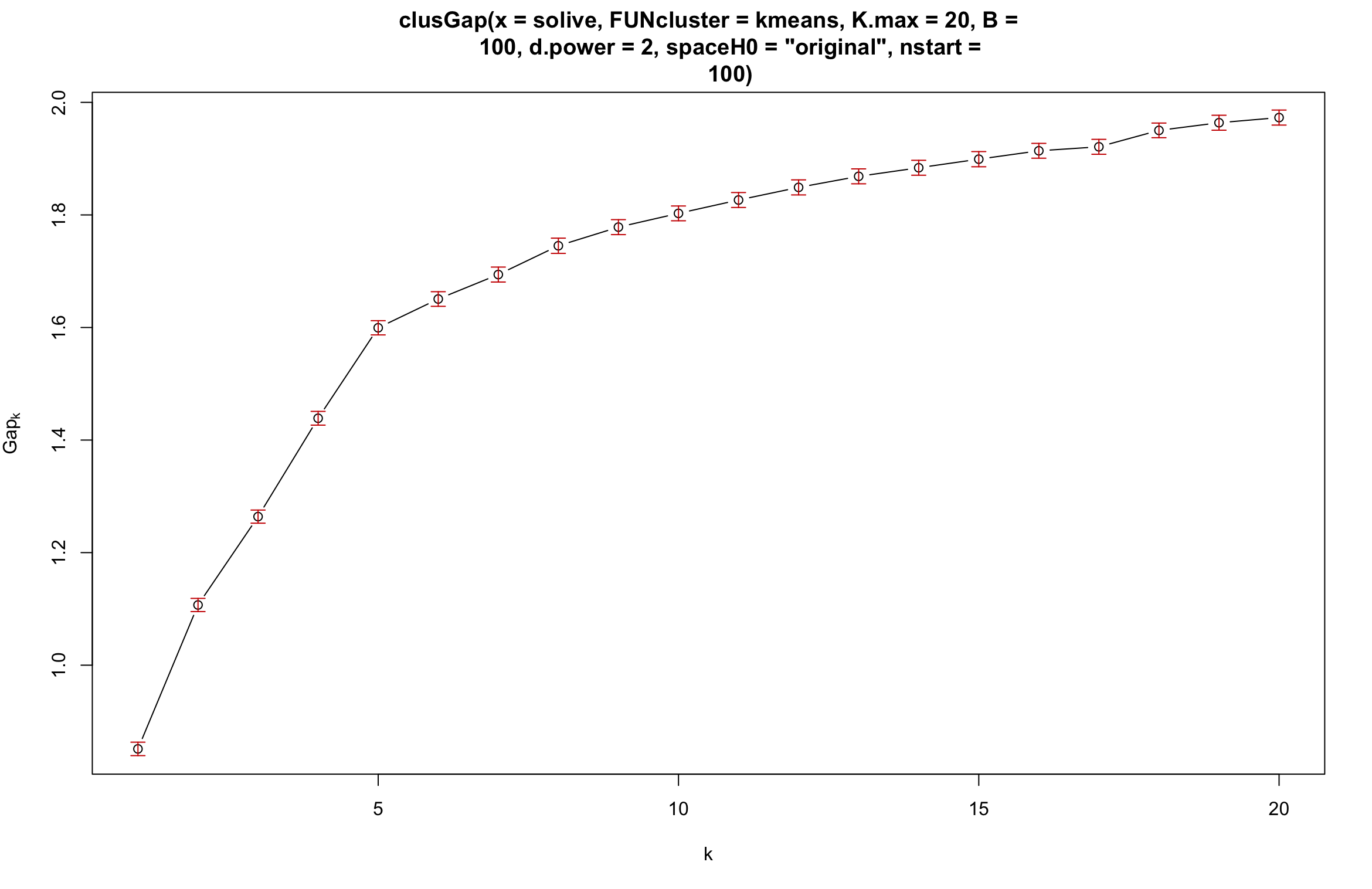




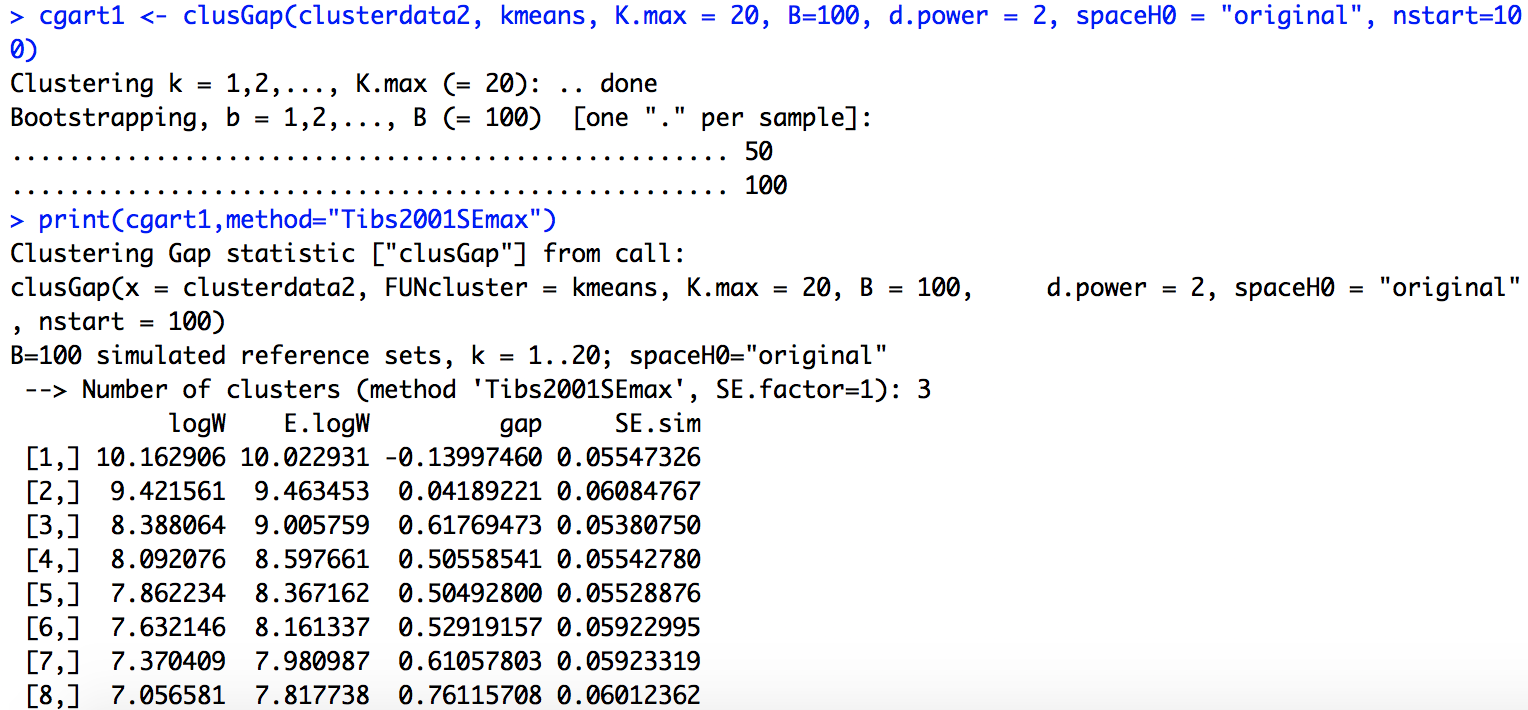
Try scaled version:

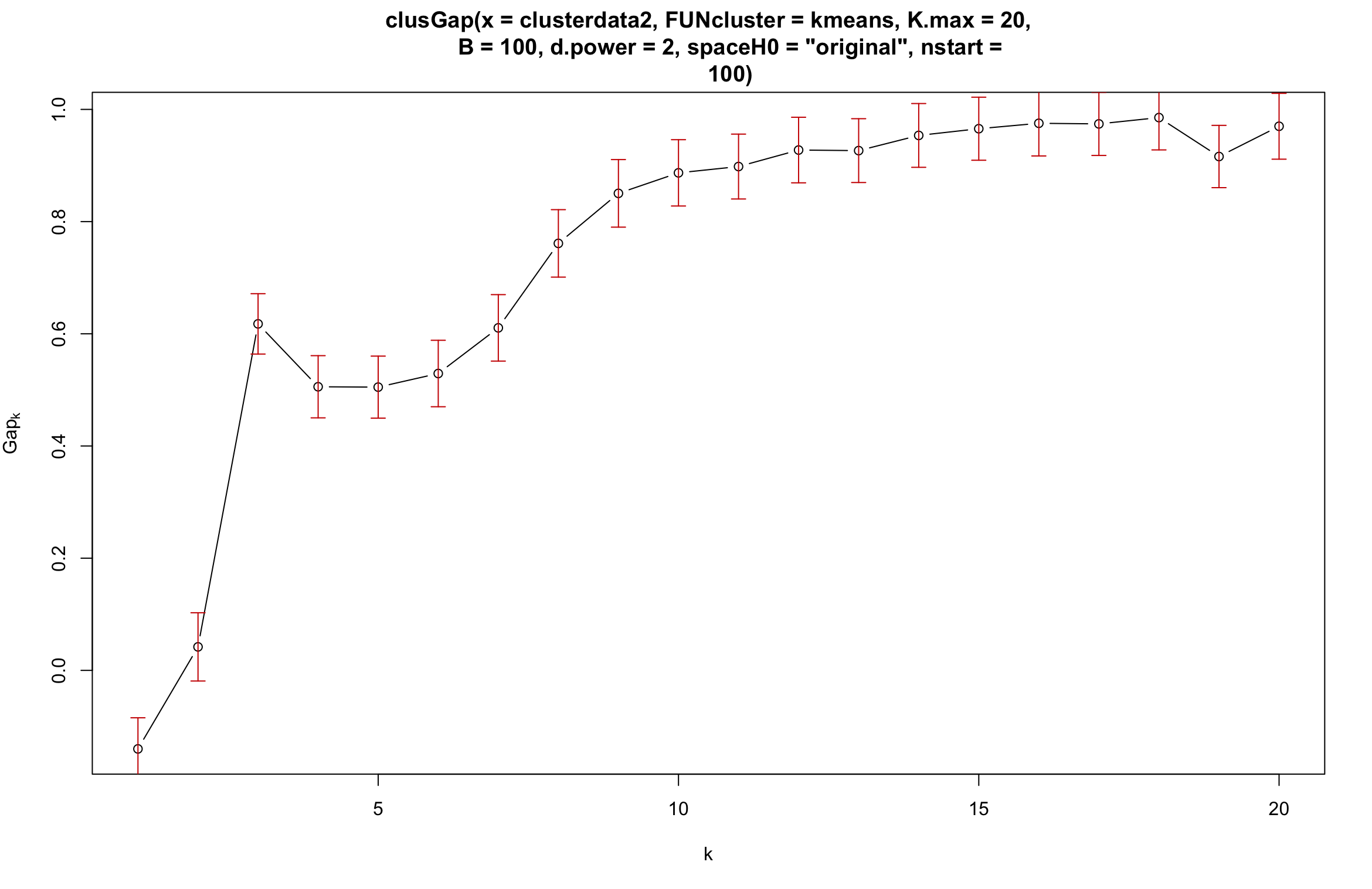






**Artificial Dataset2**





2nd time:

Clustering Gap statistic ["clusGap"] from call:

clusGap(x = clusterdata2, FUNcluster = kmeans, K.max = 20, B = 100, d.power = 2, spaceH0 = "original", nstart = 100)

B=100 simulated reference sets, k = 1..20; spaceH0="original"

--> Number of clusters (method 'Tibs2001SEmax', SE.factor=1): 3

logW E.logW gap SE.sim

[1,] 10.162906 10.029403 -0.1335028 0.05459227

[2,] 9.421561 9.469605 0.0480442 0.05835591

[3,] 8.388064 9.009728 0.6216637 0.05899336

[4,] 8.092076 8.607358 0.5152824 0.05104977

[5,] 7.862234 8.374322 0.5120884 0.05589792

[6,] 7.632146 8.170778 0.5386323 0.05540255

[7,] 7.370409 7.986016 0.6156064 0.05452030

[8,] 7.056581 7.822098 0.7655178 0.05122344

[9,] 6.819451 7.674705 0.8552532 0.05177619

[10,] 6.653562 7.544946 0.8913842 0.05169753

[11,] 6.526397 7.425600 0.8992030 0.05481559

[12,] 6.387970 7.314653 0.9266830 0.05500285

[13,] 6.257471 7.213332 0.9558614 0.05499722

[14,] 6.172711 7.117139 0.9444282 0.05268865

[15,] 6.059074 7.028204 0.9691297 0.05300277

[16,] 5.995895 6.946897 0.9510025 0.05035853

[17,] 5.924015 6.866469 0.9424542 0.04896339

[18,] 5.873572 6.793493 0.9199207 0.04896972

[19,] 5.738782 6.718450 0.9796680 0.04910523

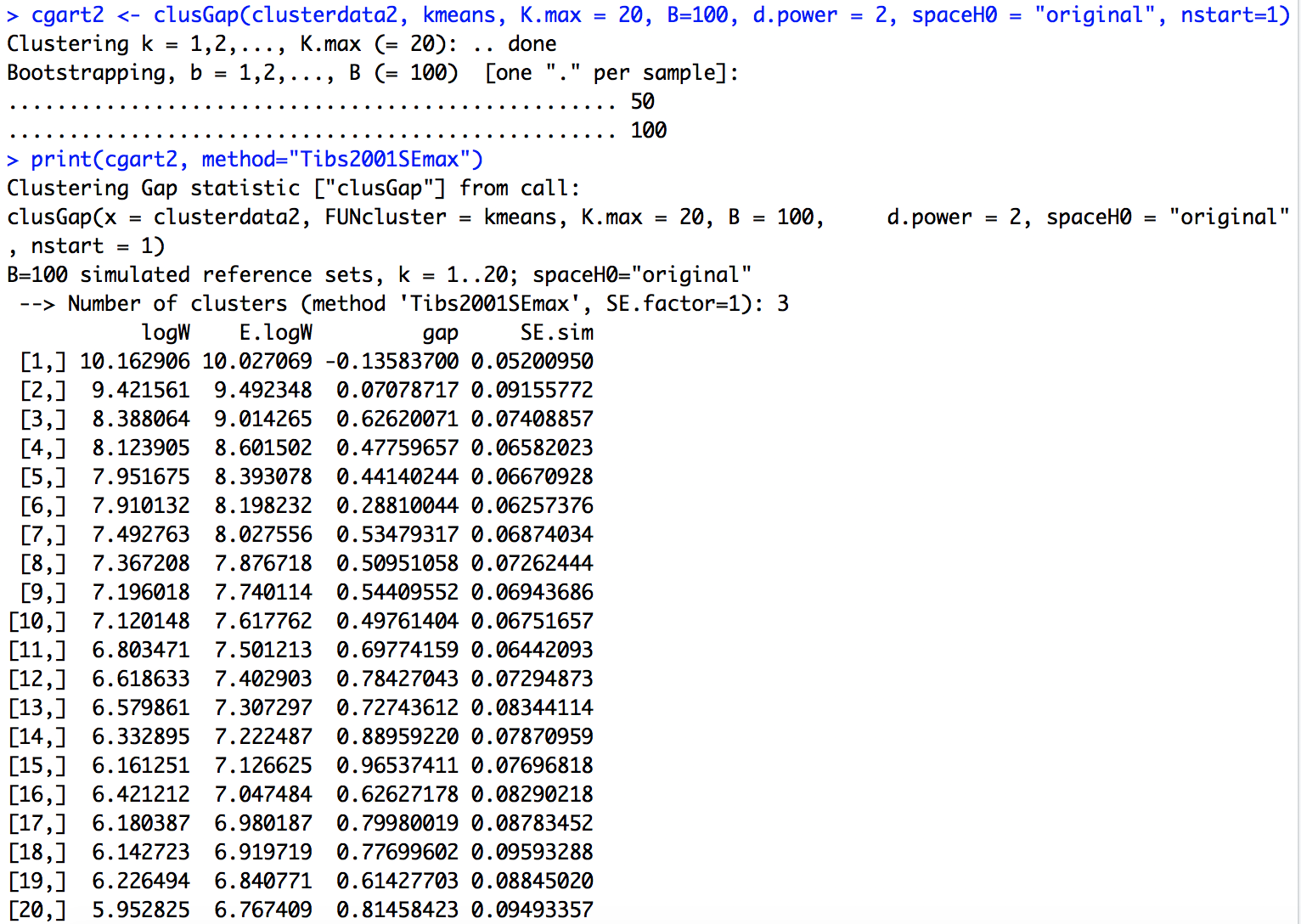
[20,] 5.717040 6.647706 0.9306669 0.04936856

Same

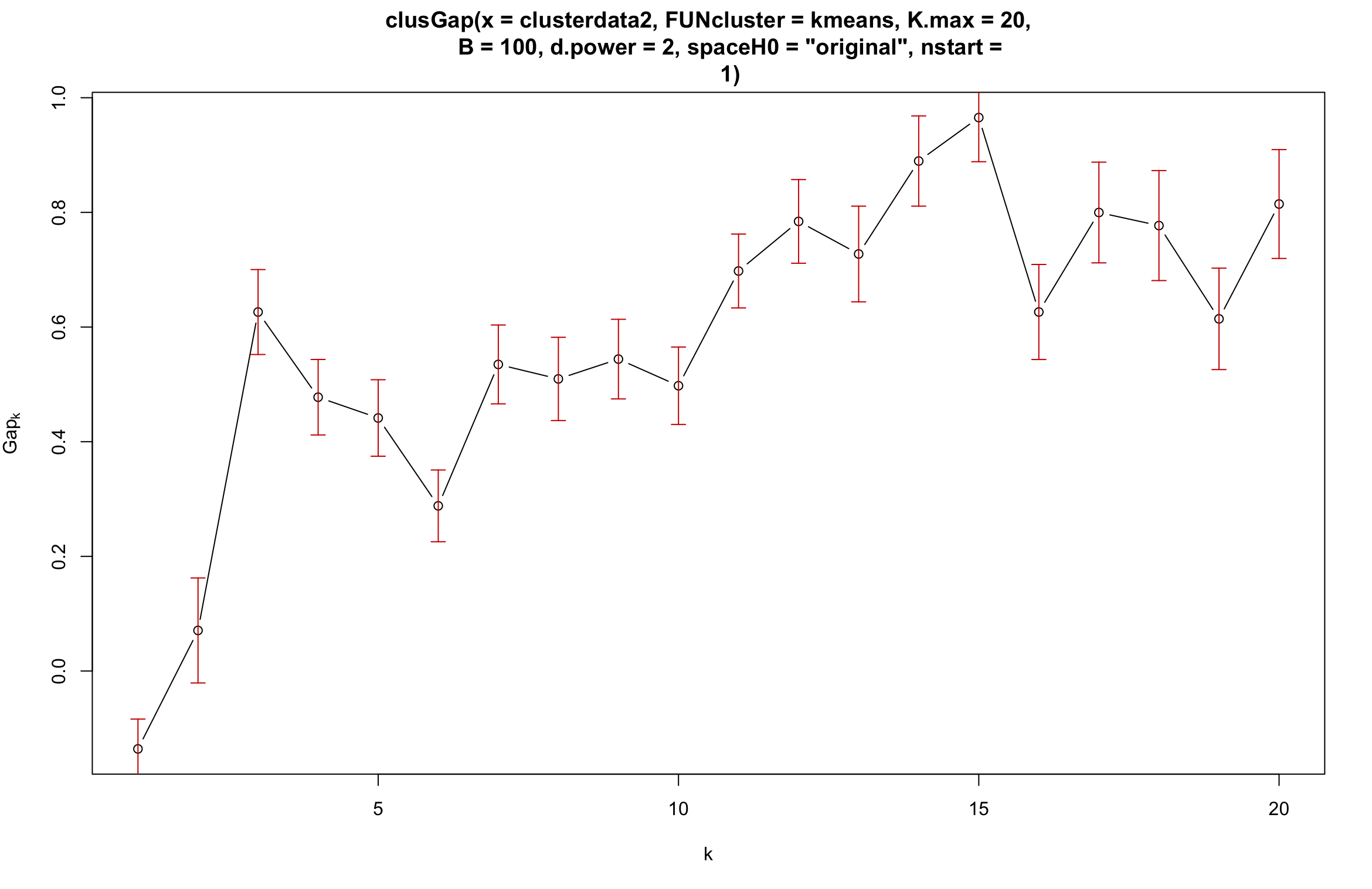
3rd time:

same

The results are the same, maybe because the nstart that I chose is large enough.



Hah! Some differences appears when I changed the nstart to 1.



Yea quite lot differences.

**Conclusion**: set a high nstart to gain a relatively stable result.

## **Question 3:**

Let’s simulate an experimental dataset with a known cluster, say 5, from 5 different distributions. And then tried the different methods to see which one performs better.

Well maybe just used the artificial dataset 2 for a try!

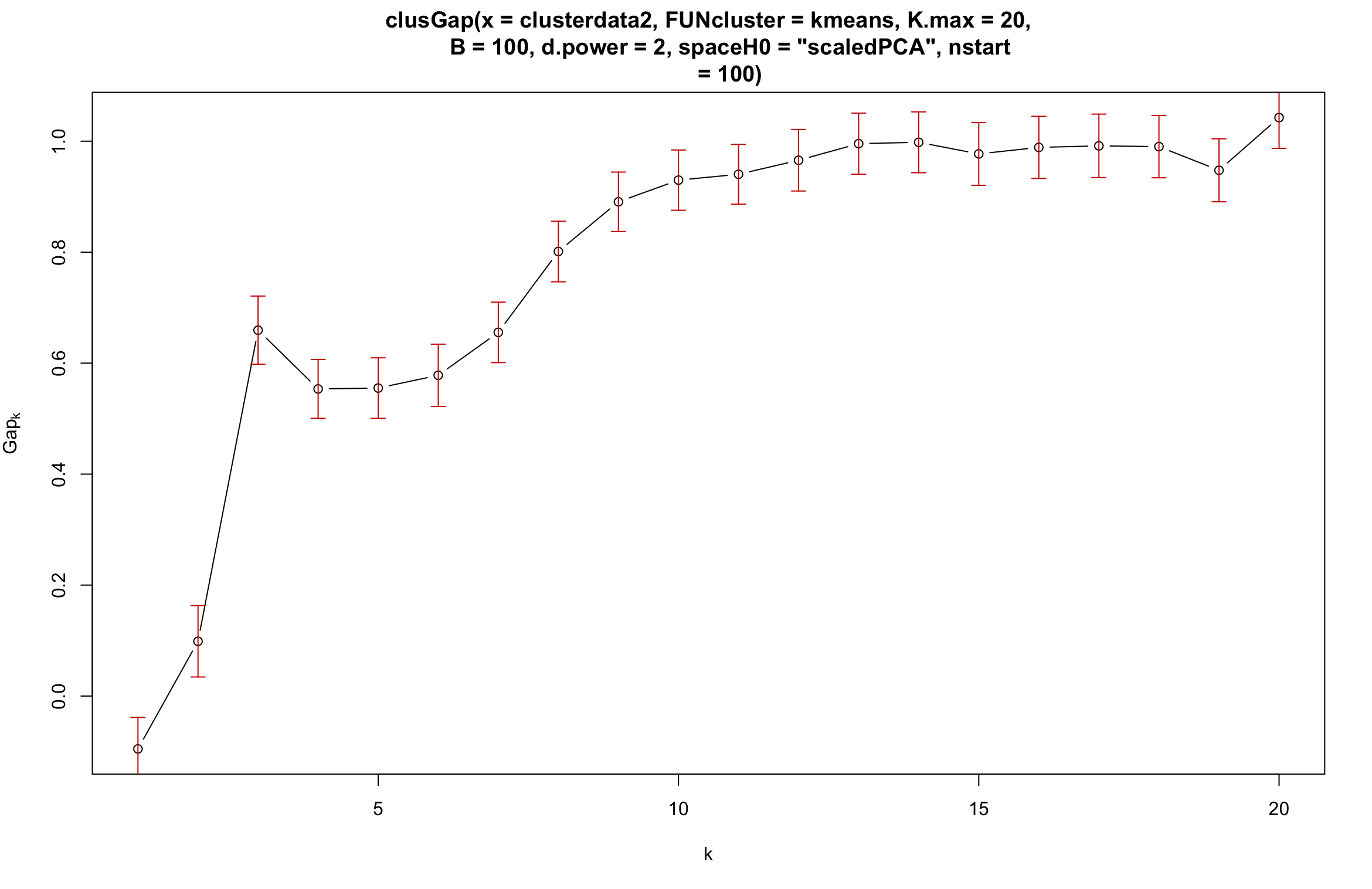
* R

set.seed(1234567)

cgart3 <- clusGap(clusterdata2, kmeans, K.max = 20, B=100, d.power = 2,spaceH0 = "scaledPCA",nstart=100)

print(cgart3,method="Tibs2001SEmax")

plot(cgart3)



There ain’t that much difference for orginal and scaledPCA.

> print(cgart1, method= "firstSEmax")

Clustering Gap statistic ["clusGap"] from call:

clusGap(x = clusterdata2, FUNcluster = kmeans, K.max = 20, B = 100, d.power = 2, spaceH0 = "original", nstart = 100)

B=100 simulated reference sets, k = 1..20; spaceH0="original"

--> Number of clusters (method 'firstSEmax', SE.factor=1): 3

logW E.logW gap SE.sim

[1,] 10.162906 10.030573 -0.13233240 0.05178483

[2,] 9.421561 9.472113 0.05055164 0.06070778

[3,] 8.388064 9.011571 0.62350699 0.04647102

[4,] 8.092076 8.596701 0.50462502 0.05406854

[5,] 7.862234 8.369249 0.50701579 0.05252120

[6,] 7.632146 8.163635 0.53148963 0.05172188

[7,] 7.370409 7.982243 0.61183364 0.05057345

[8,] 7.056581 7.814050 0.75746882 0.04523709

[9,] 6.819451 7.666732 0.84728009 0.04328101

[10,] 6.653562 7.536085 0.88252315 0.04420563

[11,] 6.526397 7.415392 0.88899506 0.04650629

[12,] 6.387970 7.305939 0.91796935 0.04750822

[13,] 6.291425 7.202226 0.91080104 0.04946357

[14,] 6.200447 7.108344 0.90789681 0.04965334

[15,] 6.097031 7.019166 0.92213499 0.05033007

[16,] 5.998996 6.933476 0.93448022 0.05316976

[17,] 5.942038 6.852122 0.91008442 0.05410203

[18,] 5.802220 6.777149 0.97492881 0.05687920

[19,] 5.800740 6.706896 0.90615646 0.05631342

[20,] 5.732909 6.637159 0.90425018 0.05933496

There are different results but the numbers of clusters found are the same.

## **Question 4:**