# chapter-clustering-validation

Ar

2022-07-19

## Need for validation of clusters

Clustering algorithms are designed such that they come out with a given number of clusters even if the underlying data is devoid of any such clusters. We will see a criterion to assess the credibility of the clusters produced by any clustering algorithm.

## Within-groups sum of squared distances (WSS):

$$WSS_k = \sum_{l=1}^{k} \sum_{x_i \in C_l} d^2(x_i, \overline{x_l})$$

1

where, $k$ is the number of clusters and within the l-th cluster $C_l$, $x_l$ is the centre of mass. We are interested in finding the *elbow* where there is a sudden drop in WSS_k as k is increased.

## Calinski-Harabasz index

There was an issue with using just the WSS_k. Sometimes, there are more than one sudden drops (Try running the example with datasets that vary in the spread from the center of the chosen cluster). For some cases, there might be more than one drop. The Calinski-Harabasz index overcomes this problem by taking into consideration the distance between the clusters as well. The formula for the CH index is as follows.

$$CH(k) = \frac{BSS_k}{WSS_k} \times \frac{N-k}{N-1} \quad where \quad BSS_k = \sum_{l=1}^{k} n_l (\bar{x_l} - \bar{x})^2$$

```
simdat
```

```
## # A tibble: 600 x 3
##         x       y class
##     <dbl>   <dbl> <chr>
##  1 -2.23  -2.05   0:0
##  2  1.66   2.15   0:0
##  3 -0.914  0.822  0:0
##  4  0.715 -1.05   0:0
##  5  0.169  1.98   0:0
##  6 -0.568 -1.56   0:0
##  7 -0.427  1.37   0:0
##  8 -0.937 -1.09   0:0
##  9  0.871 -0.936  0:0
## 10  1.11   0.0897 0:0
## # ... with 590 more rows
## # i Use `print(n = ...)` to see more rows
```

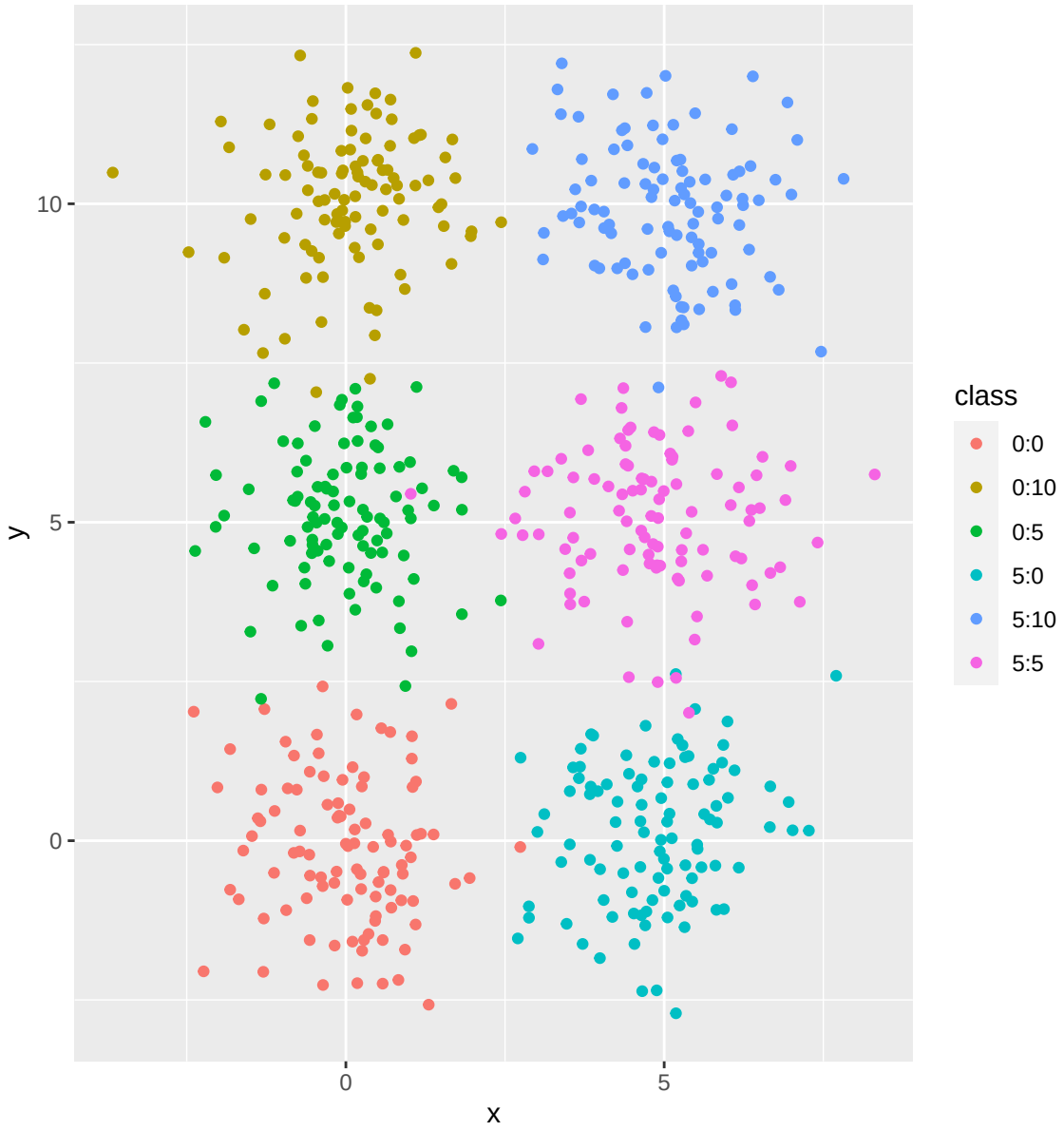


Figure 1: simulated (fake) data with six clusters.

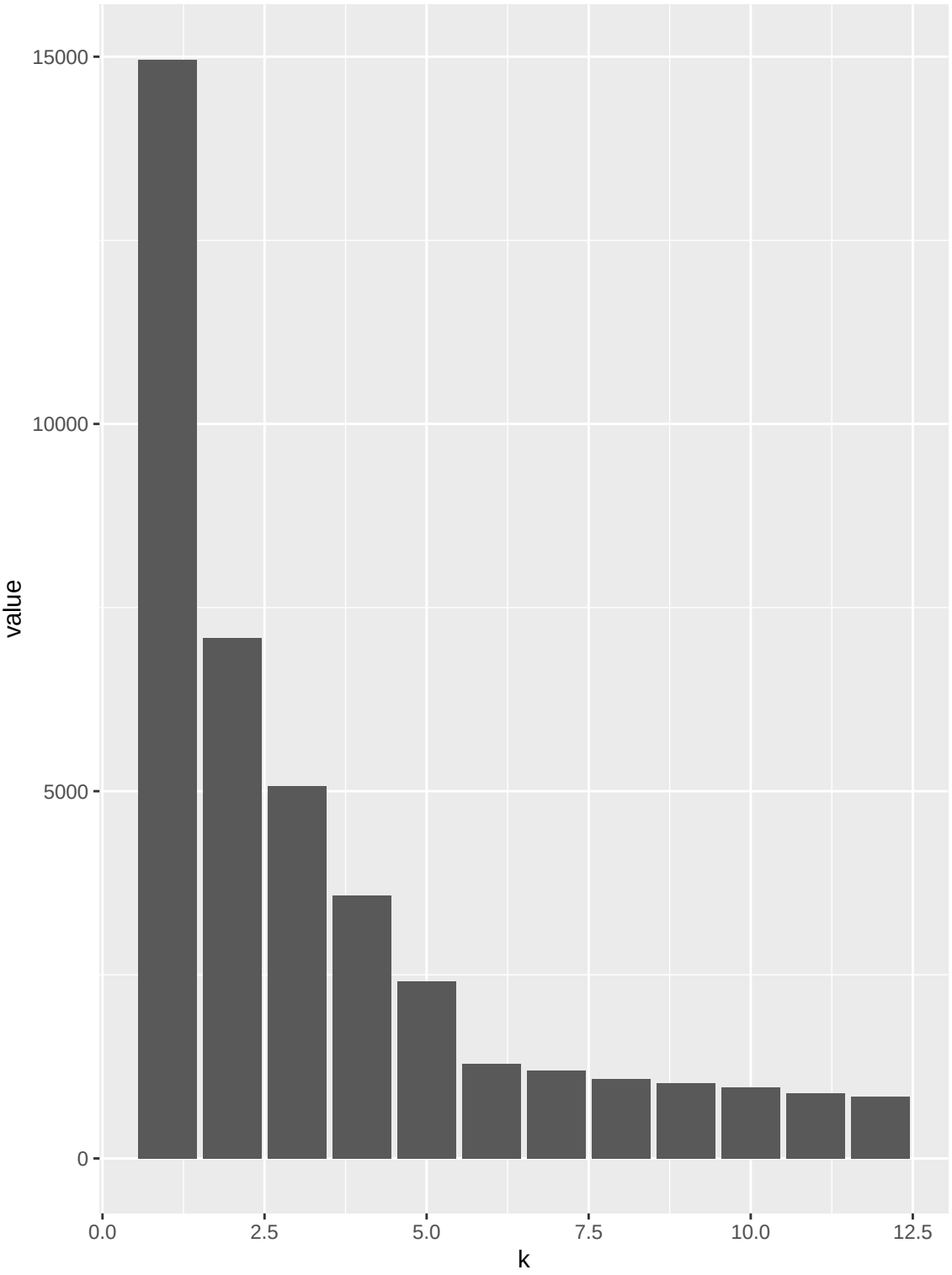


Figure 2: Reduction of within group distance with number of clusters.

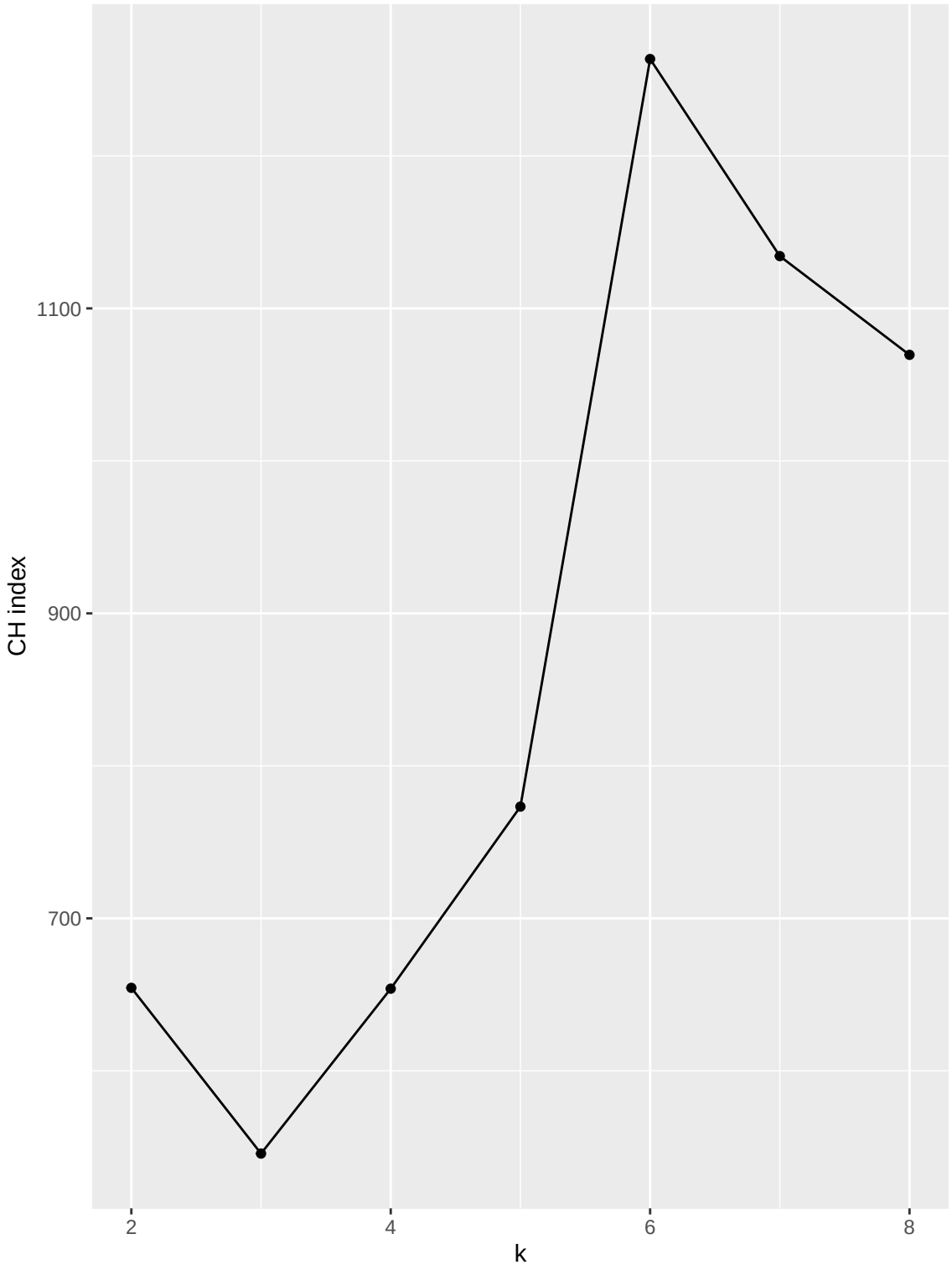


Figure 3: The peak of CH index imply an inherent structure to the data. Notice the peak when the number of clusters is six.

```
library("cluster")
```

```
pamfun = function(x, k)
  list(cluster=pam(x,k,cluster.only = TRUE))

gss=clusGap(simdatxy, FUN=pamfun, K.max=8, B=50, verbose=FALSE)

gss
```

```
## Clustering Gap statistic ["clusGap"] from call:
## clusGap(x = simdatxy, FUNcluster = pamfun, K.max = 8, B = 50,     verbose = FALSE)
## B=50 simulated reference sets, k = 1..8; spaceH0="scaledPCA"
##  --> Number of clusters (method 'firstSEmax', SE.factor=1): 2
##          logW   E.logW       gap     SE.sim
## [1,] 6.850876 6.967028 0.1161518 0.01555762
## [2,] 6.485584 6.654078 0.1684942 0.01372146
## [3,] 6.301995 6.456613 0.1546178 0.01424133
## [4,] 6.072521 6.265787 0.1932661 0.01379591
## [5,] 5.874470 6.158502 0.2840315 0.01559228
## [6,] 5.613668 6.060441 0.4467733 0.01858207
## [7,] 5.573428 5.981252 0.4078238 0.01370120
## [8,] 5.524576 5.904727 0.3801511 0.01212775
```

```
plot_gap = function(x){

  gstab = data.frame(x$Tab, k=seq_len(nrow(x$Tab)))
  ggplot(gstab, aes(k, gap))+geom_line()+
    geom_errorbar(aes(ymax=gap+SE.sim,
                      ymin=gap-SE.sim), width=0.1)+
    geom_point(size=3, col="red")
}
```

## Verifying the absence of *elbow* in data with no clusters

Here the clustering algorithm is run on a data that shares the range with the previous data (with six clusters) but has no inherent structure to it. The samples were drawn from uniform distribution.

```
simdat.unif <- lapply(c(0, 5), function(mx) {
    lapply(c(0, 5, 10), function(my) {
        tibble(
            x = runif(100, min = min(simdat$x), max = max(simdat$x)),
            y = runif(100, min = min(simdat$y), max = max(simdat$y)),
            class = "unif"
        )
    }) %>% bind_rows()
}) %>% bind_rows()
```

## Verifying the absence of *peaks* above standard-errors of neighbouring values in data with no clusters

As the Figure 7 shows, for dataset with no inherent structure, there is no peak in the graph of gap statistic as a function of the number of clusters. Contrast this with the graph shown in Figure 4

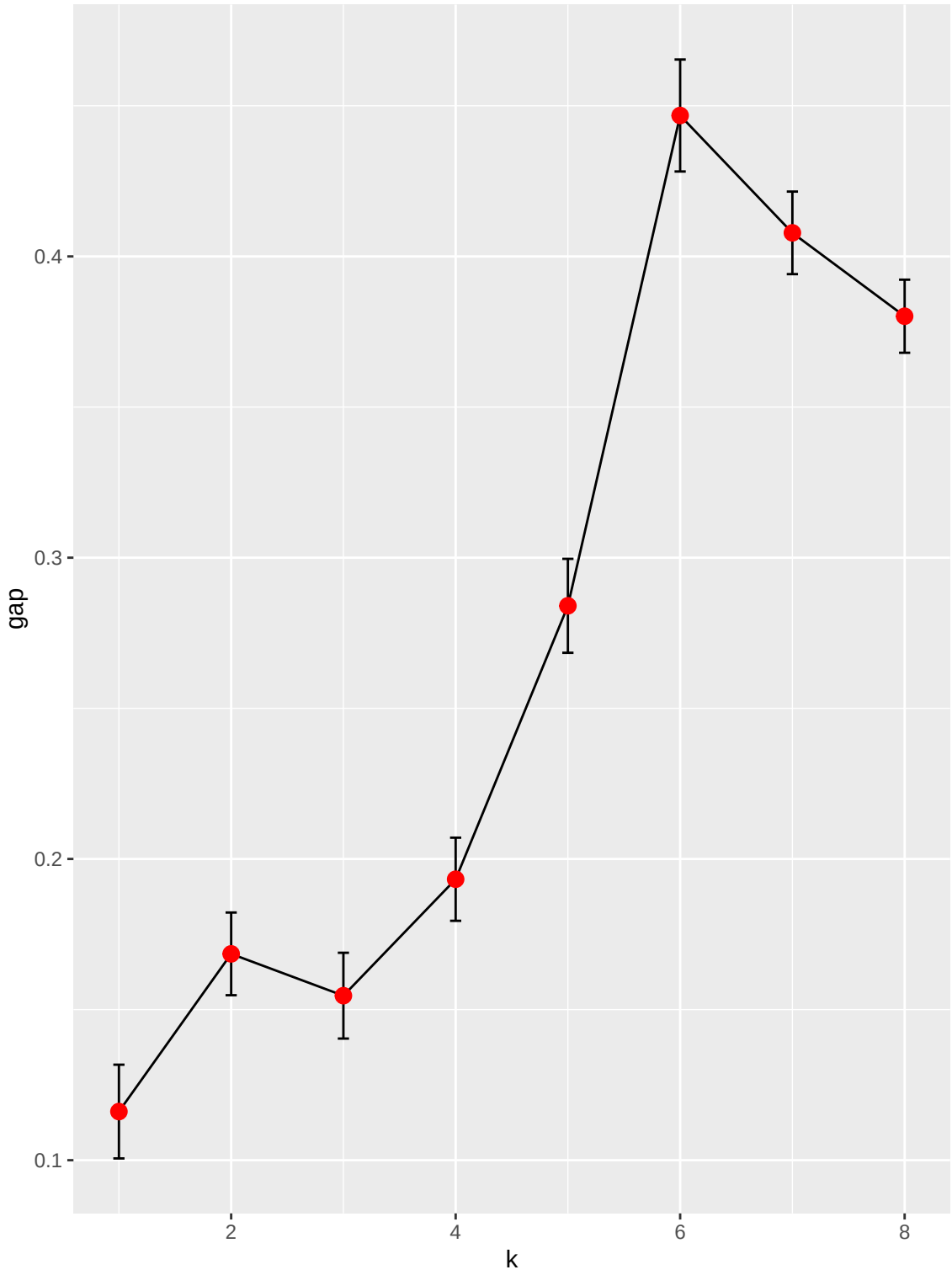


Figure 4: A peak in gap statistic also indicate a six-clustered structure.

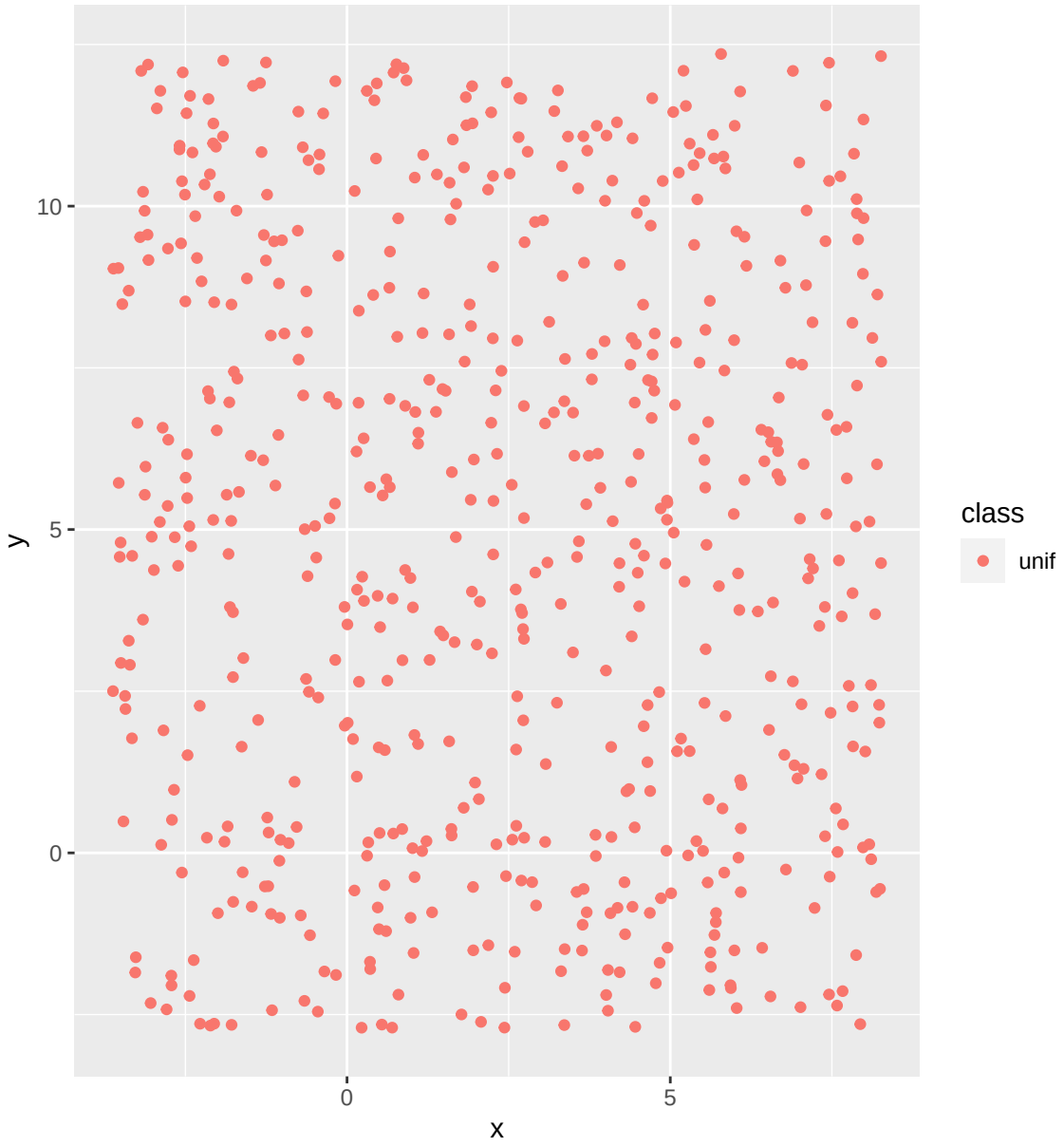


Figure 5: Simulated data with no inherent structure.

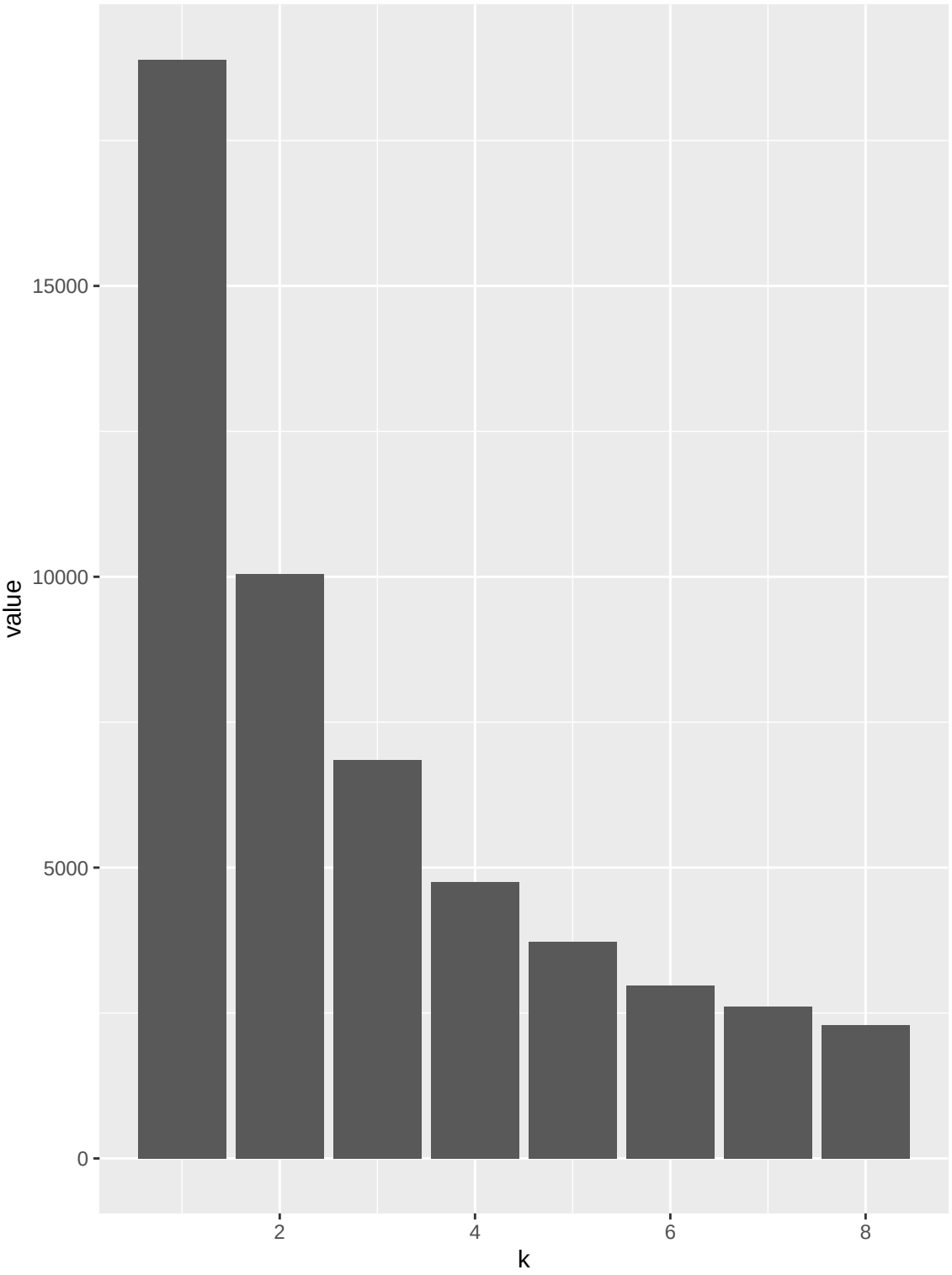


Figure 6: Reduction of within cluster summed-distances show no indication of any elbow .

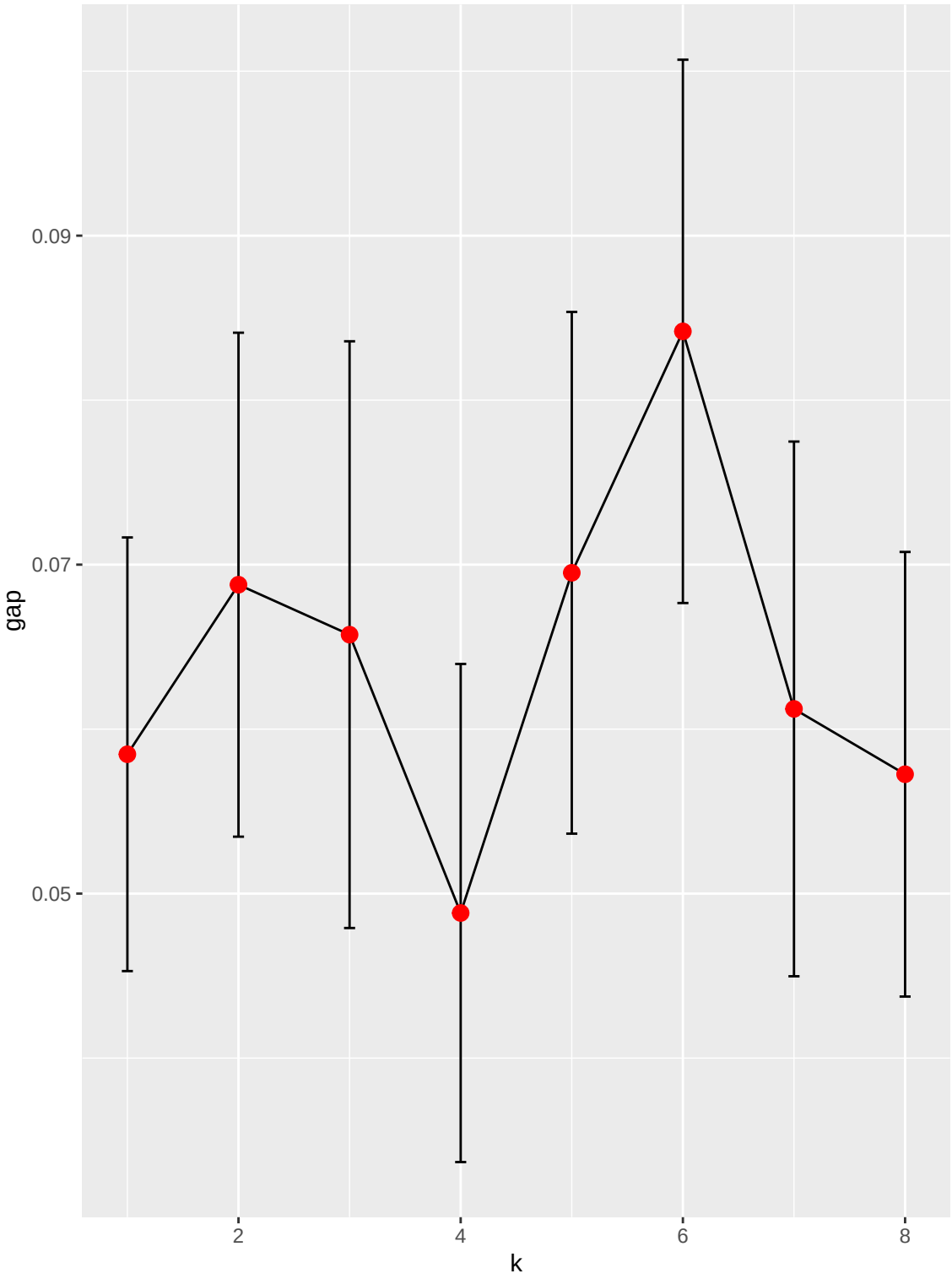


Figure 7: Absence of peak in the graph of gap statistic as a function of the number of clusters.