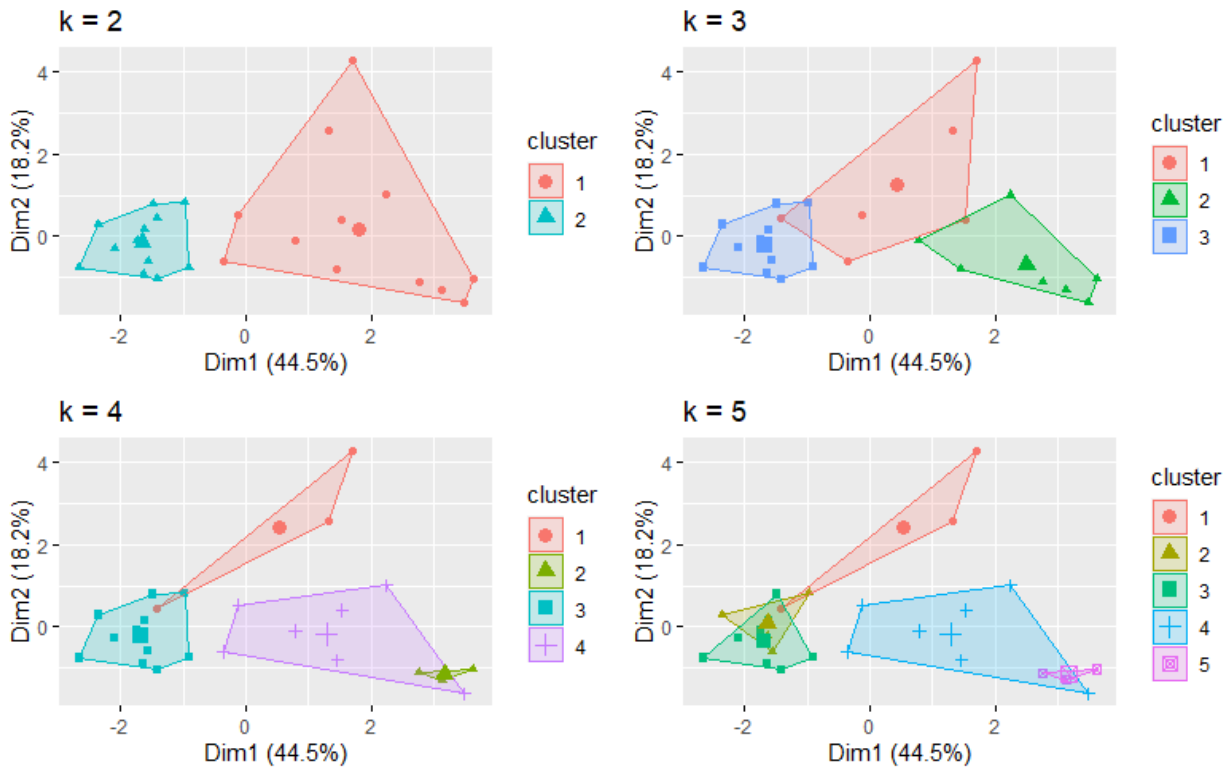
پکیج های مورد نیاز را نصب و داده‌ها را فراخوانی میکنیم :

```
> library(tidyverse) # data manipulation
> library(cluster) # clustering algorithms
> library(factoextra) # clustering algorithms & visualization
> data= read.csv("~/prot.csv")
> data1=data[,2:10]
```

حال خوشه بندی به روش **K-mens** را با ۲ تا ۵ خوشه اجرا میکنیم :

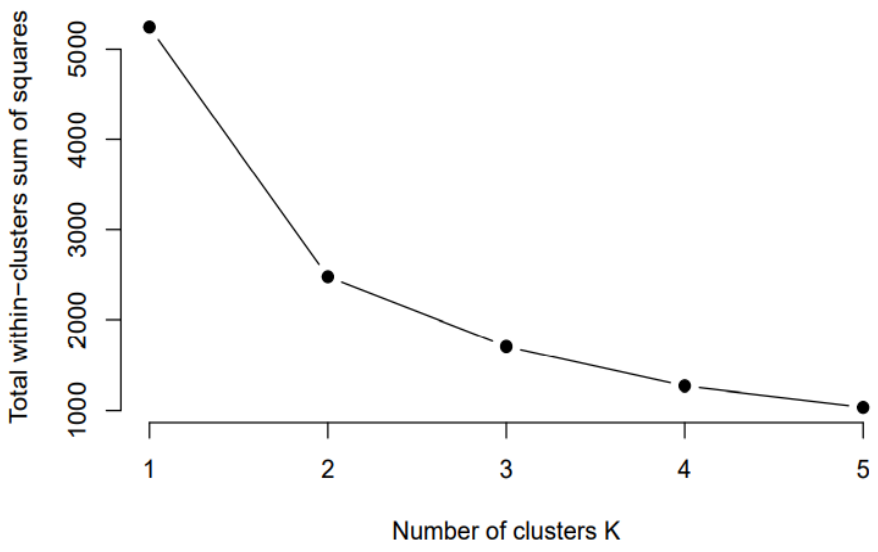
```
> k2 <- kmeans(data1, centers = 2, nstart = 25)
> k3 <- kmeans(data1, centers = 3, nstart = 25)
> k4 <- kmeans(data1, centers = 4, nstart = 25)
> k5 <- kmeans(data1, centers = 5, nstart = 25)
> # plots to compare
> p1 <- fviz_cluster(k2, geom = "point", data = data1) + ggtitle("k = 2")
> p2 <- fviz_cluster(k3, geom = "point", data = data1) + ggtitle("k = 3")
> p3 <- fviz_cluster(k4, geom = "point", data = data1) + ggtitle("k = 4")
> p4 <- fviz_cluster(k5, geom = "point", data = data1) + ggtitle("k = 5")
> #library(gridExtra)
> grid.arrange(p1, p2, p3, p4, nrow = 2)
```



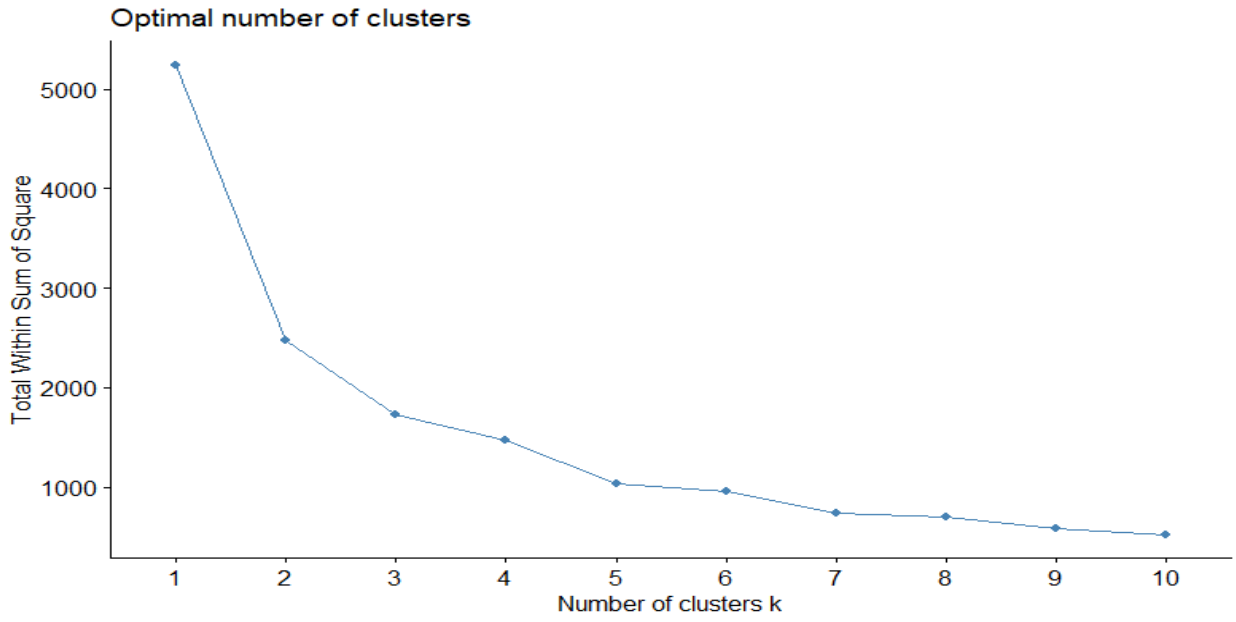
بنظر میرسد دو خوشه داده‌های ما را به خوبی جداسازی میکنند.

اکنون با روش‌های مختلف به تعیین تعداد خوشه بهینه میپردازیم :

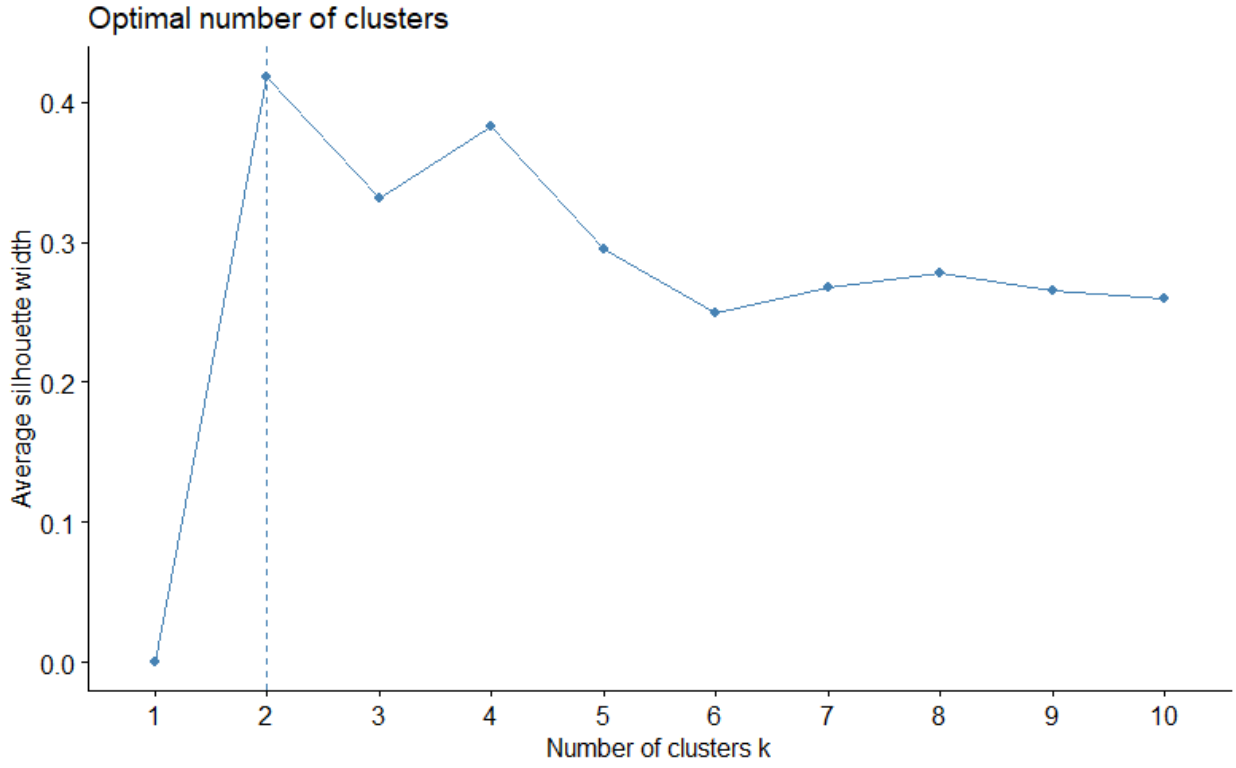
```
> #Determining Optimal Clusters
> set.seed(123)
> # function to compute total within-cluster sum of square
> wss <- function(k) {
+   kmeans(data1, k, nstart = 25 )$tot.withinss
+ }
> # Compute and plot wss for k = 1 to k = 5
> k.values <- 1:5
> # extract wss for 2-5 clusters
> wss_values <- map_dbl(k.values, wss)
> plot(k.values, wss_values,
+      type="b", pch = 19, frame = FALSE,
+      xlab="Number of clusters K",
+      ylab="Total within-clusters sum of squares")
```



```
> set.seed(123)
> fviz_nbclust(data1, kmeans, method = "wss")
```

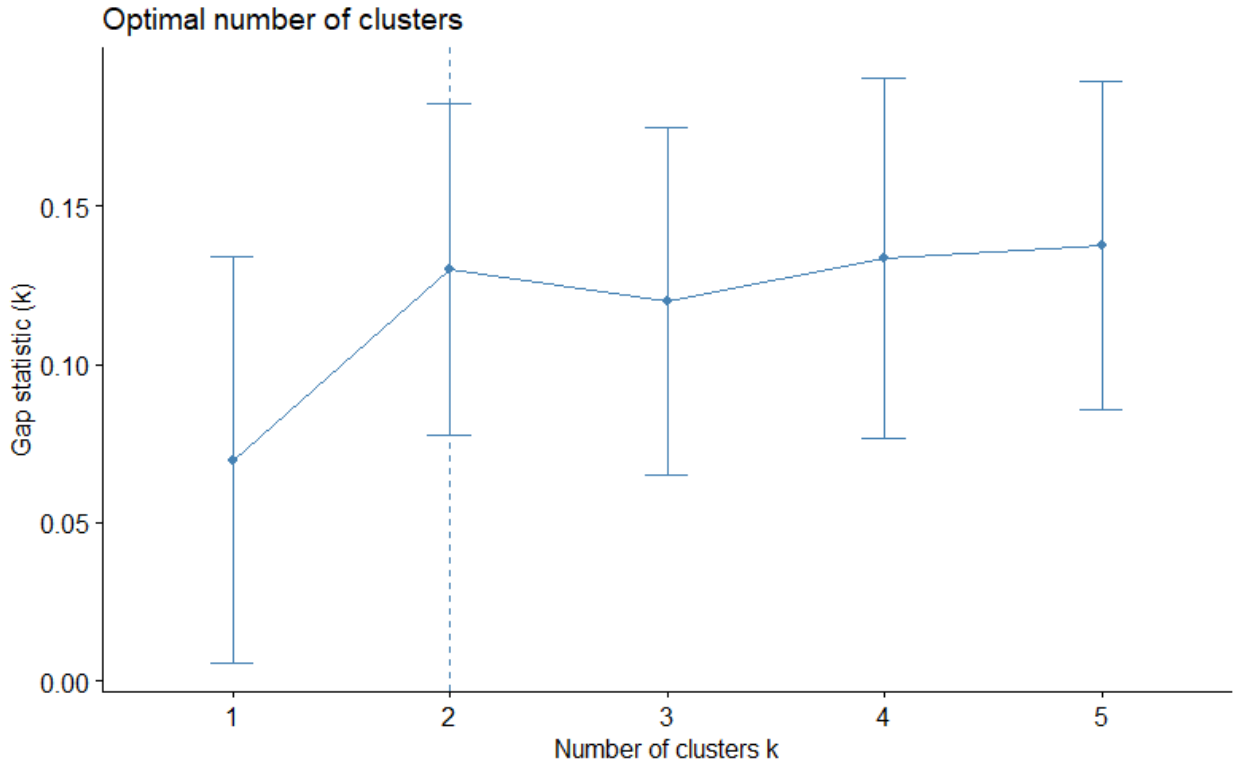



```
> #Similar to the elbow method
> fviz_nbclust(data1, kmeans, method = "silhouette")
```




```
> #Gap Statistic Method
```

```
> # compute gap statistic
> set.seed(123)
> gap_stat <- clusGap(data1, FUN = kmeans, nstart = 25,
+                     K.max = 5, B = 50)
Clustering k = 1,2,..., K.max (= 5): .. done
Bootstrapping, b = 1,2,..., B (= 50)  [one "." per sample]:
.................................................. 50
> # Print the result
> print(gap_stat, method = "firstmax")
Clustering Gap statistic ["clusGap"] from call:
clusGap(x = data1, FUNcluster = kmeans, K.max = 5, B = 50, nstart = 25)
B=50 simulated reference sets, k = 1..5; spaceH0="scaledPCA"
 --> Number of clusters (method 'firstmax'): 2
         logW   E.logW        gap     SE.sim
[1,] 4.738675 4.808432 0.06975644 0.06405369
[2,] 4.354579 4.484587 0.13000831 0.05248058
[3,] 4.177541 4.297603 0.12006227 0.05480971
[4,] 4.014242 4.147778 0.13353636 0.05693031
[5,] 3.885201 4.022753 0.13755198 0.05189354
> fviz_gap_stat(gap_stat)
```



مشاهده میشود تمام روش های فوق برای این داده‌ها تعداد ۲ خوشه را پیشنهاد میدهند که در نمودار زیر خوشه بندی با ۲ خوشه قابل مشاهده است :