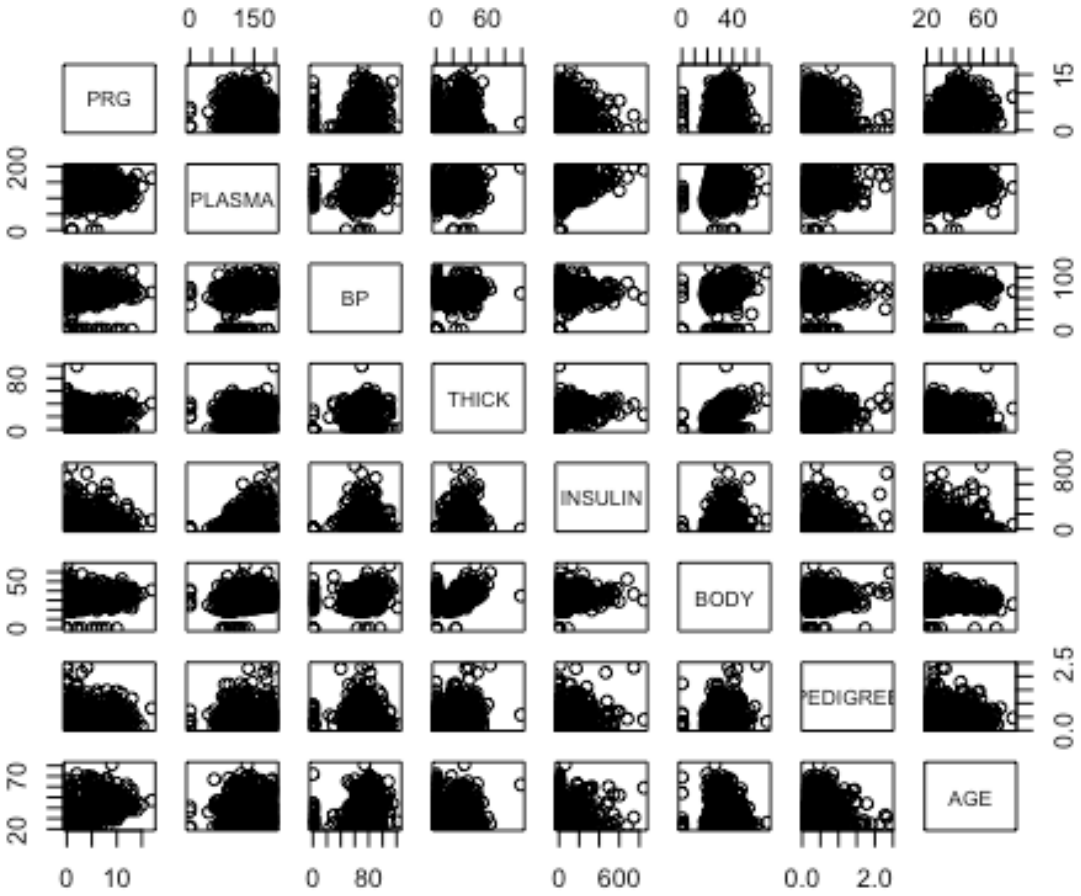
Cluster the attached Pima Indian (except the response Diabetes) data using Pam and choose the number of clusters by the 3 methods

```
> pairs(pima[,-9])
```



1. Gap statistic

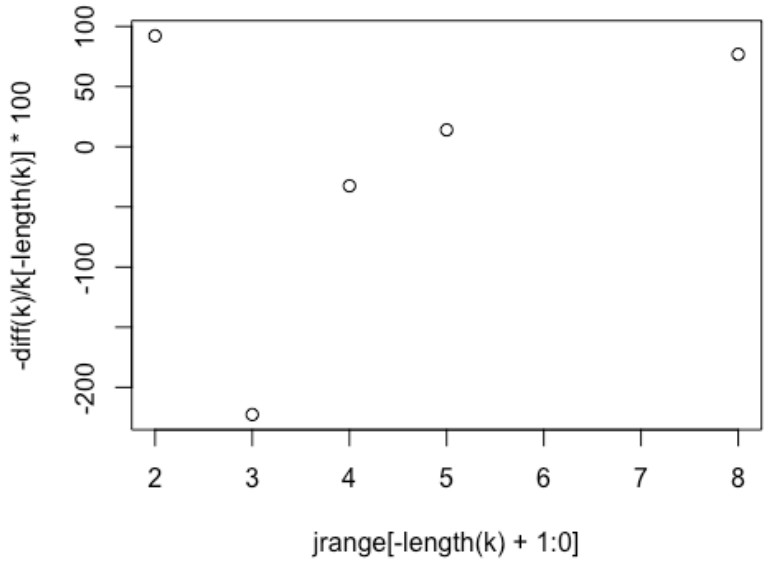
```
> x=as.matrix(pima[,c(1:8)])
> y=pima[,9]
> x0=t(t(x)/apply(x,2,sd))
> hc=hclust(dist(x0),method="ward.D")
> hc5=cutree(hc,5)
> pam1=function(x,k) list(cluster=pam(x,k,cluster.only=TRUE))
> hclust1=function(x,k)
+ list(cluster=cutree(hclust(dist(x),method="ward.D"),k))
> hk=clusGap(x0,FUN=hclust1,K.max=8,B=60)
Clustering k = 1,2,..., K.max (= 8): .. done
Bootstrapping, b = 1,2,..., B (= 60)  [one "." per sample]:
.................................................. 50
.......... 60
```

```
> hk
Clustering Gap statistic ["clusGap"] from call:
clusGap(x = x0, FUNcluster = hclust1, K.max = 8, B = 60)
B=60 simulated reference sets, k = 1..8; spaceH0="scaledPCA"
 --> Number of clusters (method 'firstSEmax', SE.factor=1): 1
       logW   E.logW       gap      SE.sim
[1,] 6.580002 7.376406 0.7964036 0.005704312
[2,] 6.494643 7.283333 0.7886896 0.007768773
[3,] 6.424285 7.248872 0.8245874 0.007135468
[4,] 6.374673 7.218607 0.8439343 0.006716839
[5,] 6.334496 7.192093 0.8575969 0.006040859
[6,] 6.301690 7.168744 0.8670546 0.006089523
[7,] 6.273943 7.147803 0.8738605 0.006441902
[8,] 6.248508 7.128665 0.8801578 0.006724834
> hkpam=clusGap(x0,FUN=pam1,K.max=8,B=60)
Clustering k = 1,2,..., K.max (= 8): .. done
Bootstrapping, b = 1,2,..., B (= 60)  [one "." per sample]:
.................................................. 50
.......... 60
> hkpam
Clustering Gap statistic ["clusGap"] from call:
clusGap(x = x0, FUNcluster = pam1, K.max = 8, B = 60)
B=60 simulated reference sets, k = 1..8; spaceH0="scaledPCA"
 --> Number of clusters (method 'firstSEmax', SE.factor=1): 7
       logW   E.logW       gap      SE.sim
[1,] 6.580002 7.377095 0.7970931 0.006491897
[2,] 6.488124 7.288262 0.8001380 0.018471507
[3,] 6.415164 7.242845 0.8276807 0.008323678
[4,] 6.366325 7.207342 0.8410173 0.009310125
[5,] 6.330204 7.179211 0.8490072 0.007743040
[6,] 6.277732 7.154127 0.8763950 0.007524557
[7,] 6.246148 7.131958 0.8858100 0.008697449
[8,] 6.229427 7.111024 0.8815973 0.008111623
```

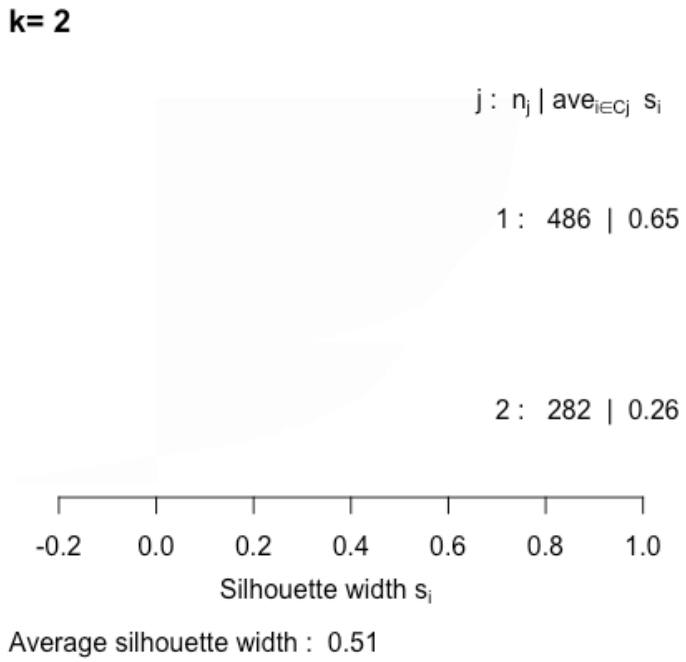
2. Second derivative

```
> clust_sel=function(x,hc,jrange=2:8,dd=2) {
+    wss4=function(x,hc,w=rep(1,length(hc))) sum(lm(x~factor(hc),weights=w)$resid^2*w)
+    sm1=NULL
+    for(i in jrange) sm1[i]=wss4(x,cutree(hc,i))
+    sm1=sm1[jrange]
+    k=if(dd==1) sm[-1] else -diff(sm1)
+    plot(jrange[-length(k)+1:0],-diff(k)/k[-length(k)]*100)
+    jrange[sort.list(diff(k)/k[-length(k)]*100)[1:4]]
+ }
```

> clust_sel(x,hc)
[1] 2 6 5 4



3. Silhouette statistic
> plot(silhouette(pam(x,k=2)),main=paste("k=",2),do.n.k=FALSE)



and compare the results.

Ans: 1) Gap statistic = 7; 2) second derivative = 2; 3) silhouette statistic = 2