K-medoides (PAM)

Karina Itzel Rodríguez Conde

27/5/2022

MÉTODO PAM

Introducción

k-medoides, también conocido como *Partitioning Around Medoids* y es un algoritmo de agrupamiento que trabaja con particiones e intenta minimizar la distancia entre puntos que se añadirían a un grupo y otro punto que es el centro de ese grupo, además de que se consideran las medianas. Divide los datos conformados por n objetos en **k** grupos. Una de sus desventajas es que no es tan eficaz para agrupar un gran conjunto de datos.

Matriz de datos

Se trabajó con la matriz **state.x77** la cual está precargada en R y contiene los 50 estados de los Estados Unidos de América. Contando con 50 filas y 8 columnas.

Exploración de la matriz

```
X<-as.data.frame(state.x77)
```

1.- Dimensión

```
dim (X)
```

```
## [1] 50  8
```

Esta base de datos contiene 50 observaciones y 8 variables.

2.- Tipos de variables

```
str(X)
```

```
## 'data.frame':    50 obs. of  8 variables:
##  $ Population: num  3615 365 2212 2110 21198 ...
##  $ Income    : num  3624 6315 4530 3378 5114 ...
##  $ Illiteracy: num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
##  $ Life Exp  : num  69 69.3 70.5 70.7 71.7 ...
```

```
##  $ Murder    : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
##  $ HS Grad   : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...
##  $ Frost     : num  20 152 15 65 20 166 139 103 11 60 ...
##  $ Area      : num  50708 566432 113417 51945 156361 ...
```

3.- Nombre de las variables

```
colnames(X)
```

```
## [1] "Population" "Income"     "Illiteracy" "Life Exp"   "Murder"
## [6] "HS Grad"    "Frost"      "Area"
```

4.- Saber si la base presenta NA

```
anyNA(X)
```

```
## [1] FALSE
```

Esta base de datos no presenta datos nulos.

Transformación de la matriz

Tratamiento de la matriz

1.- Transformación de las variables x1,x3 y x8 con la función de logaritmo.

```
X[,1]<-log(X[,1])
colnames(X)[1]<-"Log-Population"

X[,3]<-log(X[,3])
colnames(X)[3]<-"Log-Illiteracy"

X[,8]<-log(X[,8])
colnames(X)[8]<-"Log-Area"
```

Método PAM

1.- Separación de filas y columnas.

```
dim(X)
```

```
## [1] 50  8
```

```
n<-dim(X)[1]
p<-dim(X)[2]
```

2.- Estandarización univariante.

```r
X.s<-scale(X)
```

```r
library(cluster)
```

3.- Aplicación del algoritmo (3 grupos)

```r
pam.3<-pam(X.s,3)
```

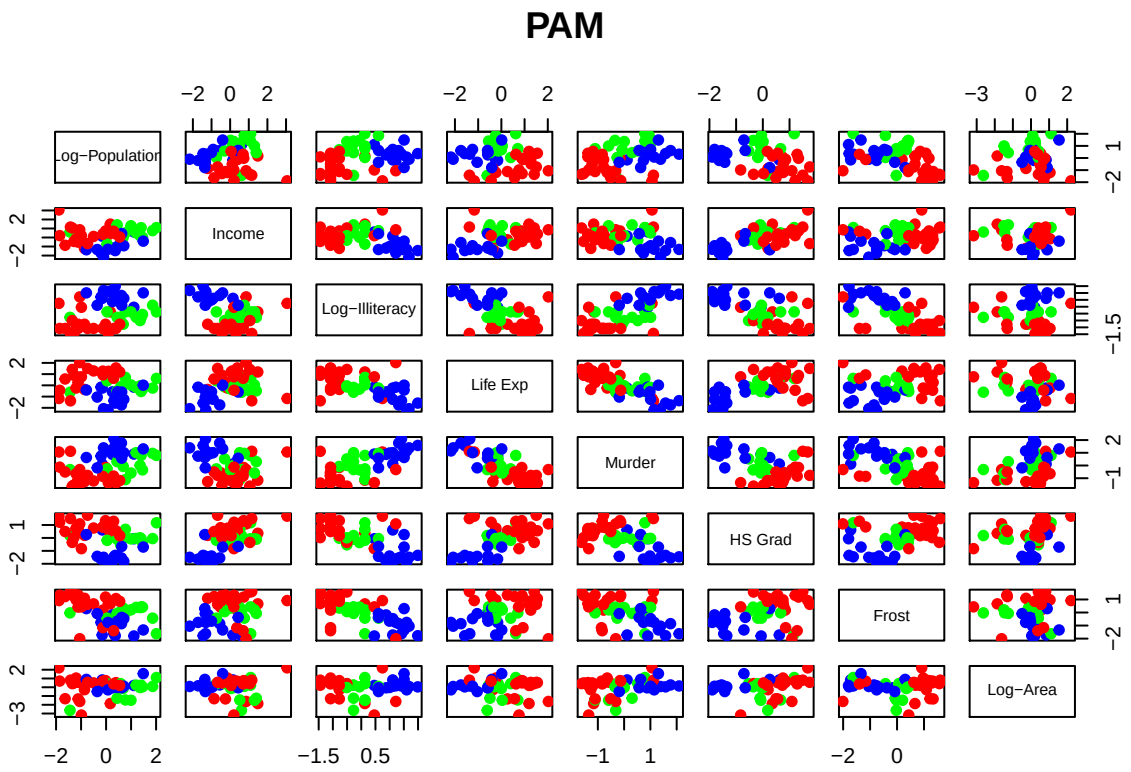
4.- Clústers

```r
cl.pam<-pam.3$clustering
cl.pam
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              2              1              1              3
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              2              3              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              2              3              3              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              1              1              2              3
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              3              2              1              3
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              2              2              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              3              1              2              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              2              3              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              1              1              2              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              2              1              2              2
```

5.- Scatter plot de la matriz con los grupos

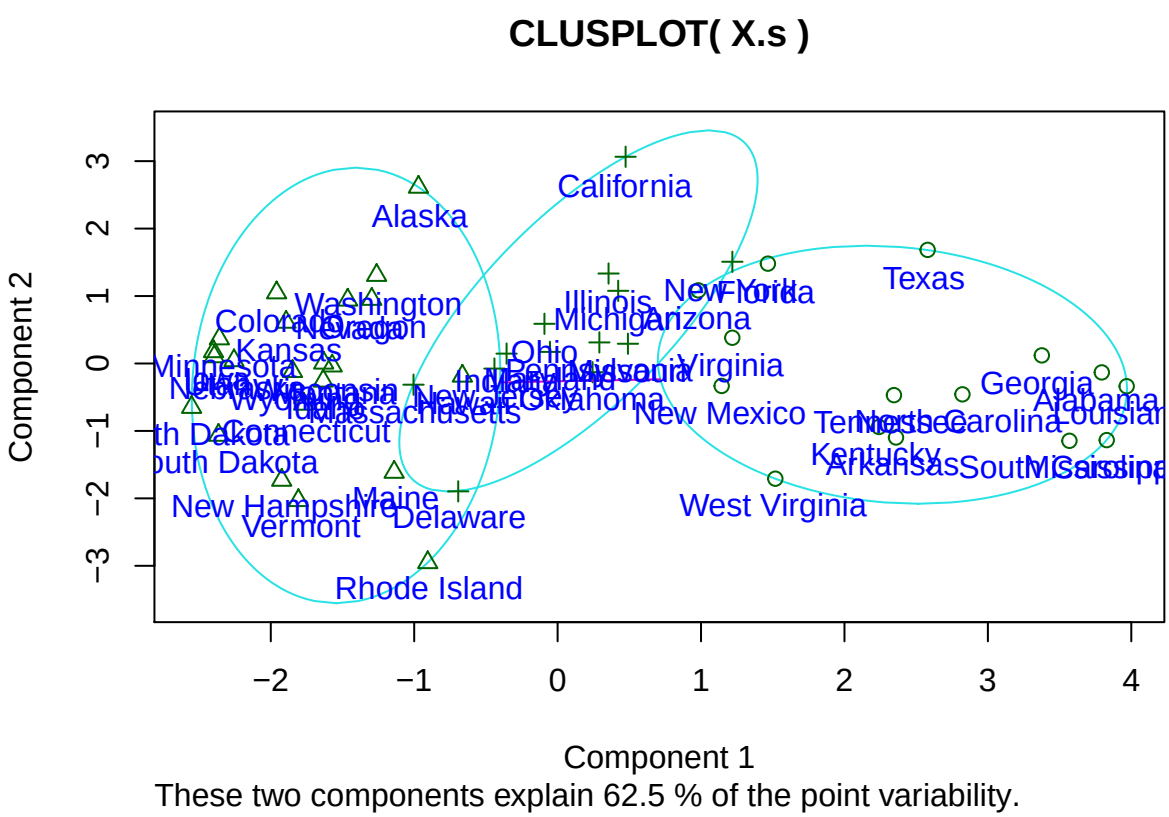
```r
col.cluster<-c("blue","red","green")[cl.pam]
pairs(X.s, col=col.cluster, main="PAM", pch=19)
```



Visualizacion con Componentes Principales

```
library(cluster)
```

```
clusplot(X.s,cl.pam)
text(princomp(X.s)$scores[,1:2],
     labels=rownames(X.s),pos=1, col="blue")
```



These two components explain 62.5 % of the point variability.

Silhouette

Representación gráfica de la eficacia de clasificación de una observación dentro de un grupo:

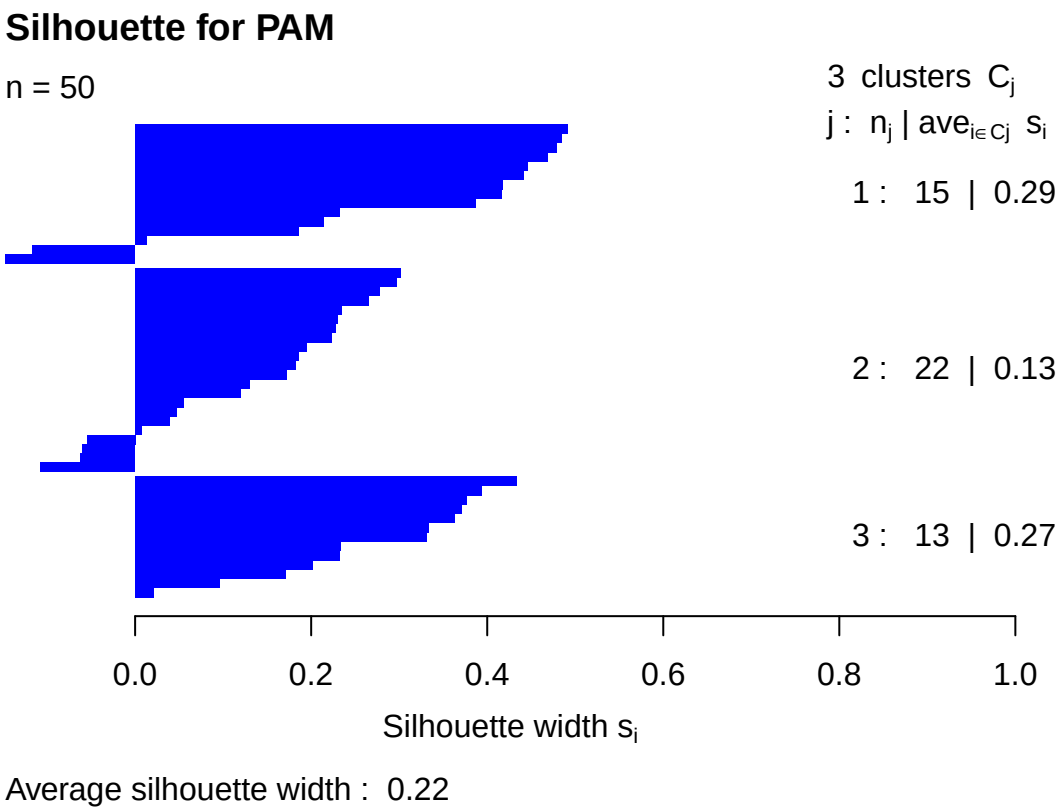
Mayor a 0.7 = mejor clasificación. Entre más cercana a 1 es mejor.

1.- Generación de los cálculos

```
dist.Euc<-dist(X.s, method = "euclidean")
Sil.pam<-silhouette(cl.pam, dist.Euc)
```

2.- Generación del gráfico

```
plot(Sil.pam, main="Silhouette for PAM",
     col="blue")
```



Dado el gráfico anterior: el clúster 1 contiene 15 estados y una probabilidad de Silhouette del 0.29, considerada como un poco baja. El clúster 2 contiene 22 estados y su probabilidad de Silhouette es del 0.13, comparado con el clúster 1, su probabilidad es más baja. Mientras que, el clúster 3 contiene 13 estados y su probabilidad es del 0.27, considerado como un poco bajo. Hay algunos datos que no se clasifican, son negativos y como el valor del Silhouette es de 0.22, es muy bajo; por lo que se necesita un valor más alto.

Debido a ello, como ejercicio, ahora se realizan 2 y 4 clústers para tomar la mejor decisión de agrupamiento.

2 Clústers

1.- Aplicacion del algoritmo (2 grupos)

```
pam.2<-pam(X.s,2)
```

2.- Clústers

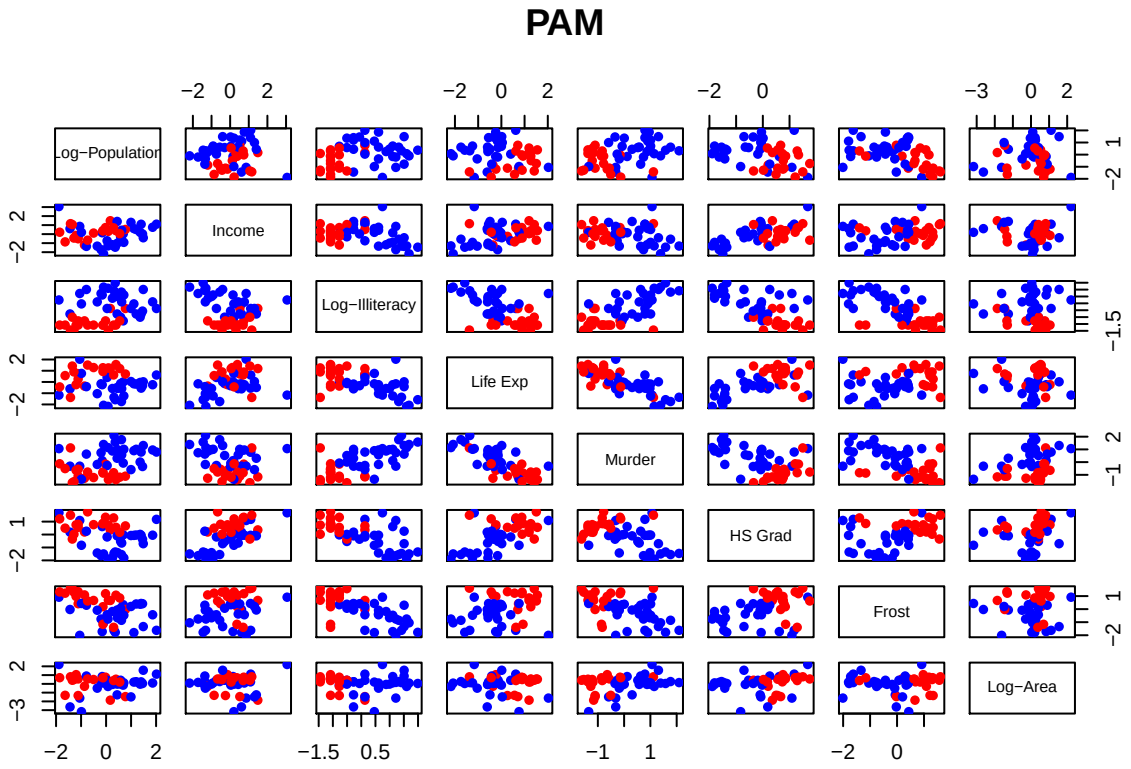
```
cl.pam<-pam.2$clustering
cl.pam
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              1              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              2              1              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              2              1              1              2
```

```
##          Kansas        Kentucky       Louisiana           Maine        Maryland
##               2               1               1               2               1
##   Massachusetts        Michigan       Minnesota     Mississippi        Missouri
##               2               1               2               1               1
##         Montana        Nebraska          Nevada   New Hampshire      New Jersey
##               2               2               2               2               1
##      New Mexico        New York  North Carolina    North Dakota            Ohio
##               1               1               1               2               1
##        Oklahoma          Oregon    Pennsylvania    Rhode Island  South Carolina
##               1               2               1               1               1
##    South Dakota       Tennessee           Texas            Utah         Vermont
##               2               1               1               2               2
##        Virginia      Washington   West Virginia       Wisconsin         Wyoming
##               1               2               1               2               2
```

3.- Scatter plot de la matriz con los grupos

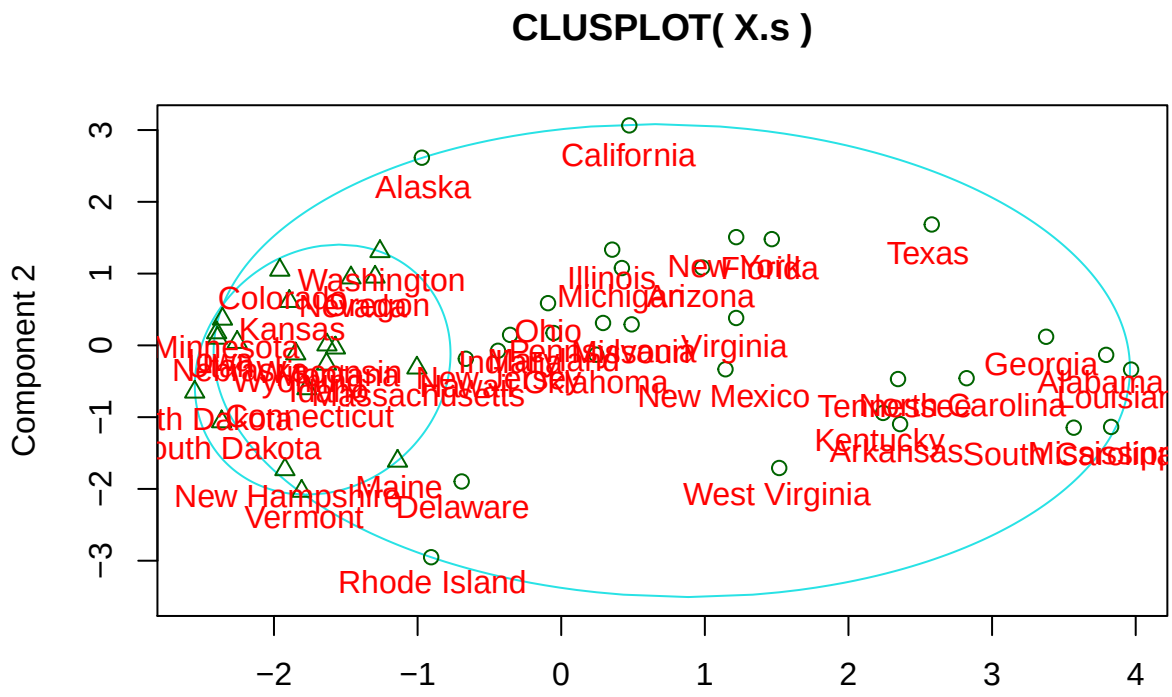
```
col.cluster<-c("blue","red")[cl.pam]
pairs(X.s, col=col.cluster, main="PAM", pch=16)
```



Visualizacion con Componentes Principales

```
library(cluster)
```

```
clusplot(X.s,cl.pam)
text(princomp(X.s)$scores[,1:2],
     labels=rownames(X.s),pos=1, col="red")
```



These two components explain 62.5 % of the point variability.

Silhouette

Representación gráfica de la eficacia de clasificación de una observación dentro de un grupo:

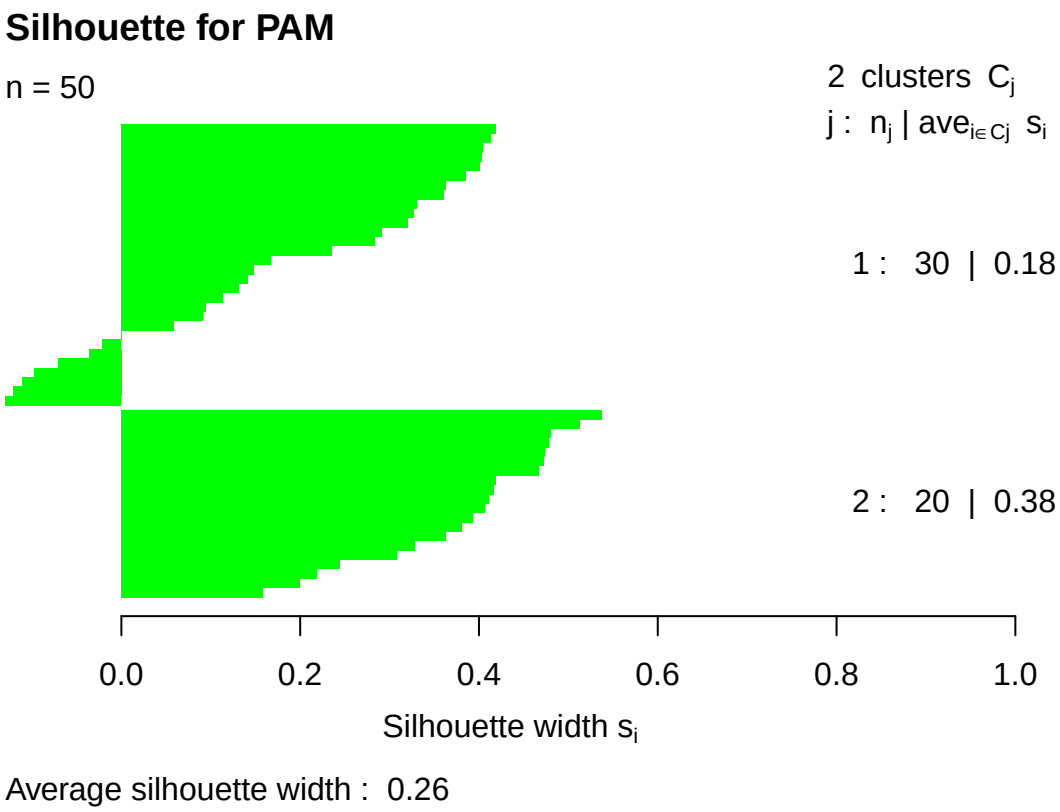
Mayor a 0.7 = mejor clasificación. Entre más cercana a 1 es mejor.

1.- Generación de los cálculos

```
dist.Euc<-dist(X.s, method = "euclidean")
Sil.pam<-silhouette(cl.pam, dist.Euc)
```

2.- Generación del gráfico

```
plot(Sil.pam, main="Silhouette for PAM",
     col="green")
```



Con base al gráfico anterior, con 2 clústers, todavía hay datos no clasificados, pero el Average del Silhouette es de 0.26. Hay un clúster con una probabilidad del Silhouette algo buena (0.38, clúster 2) y otra con probabilidad algo baja (0.18, clúster 1).

Se probará ahora con 4 clúster.

4 Clústers

1.- Aplicacion del algoritmo (4 grupos)

```
pam.4<-pam(X.s,4)
```

2.- Clústers

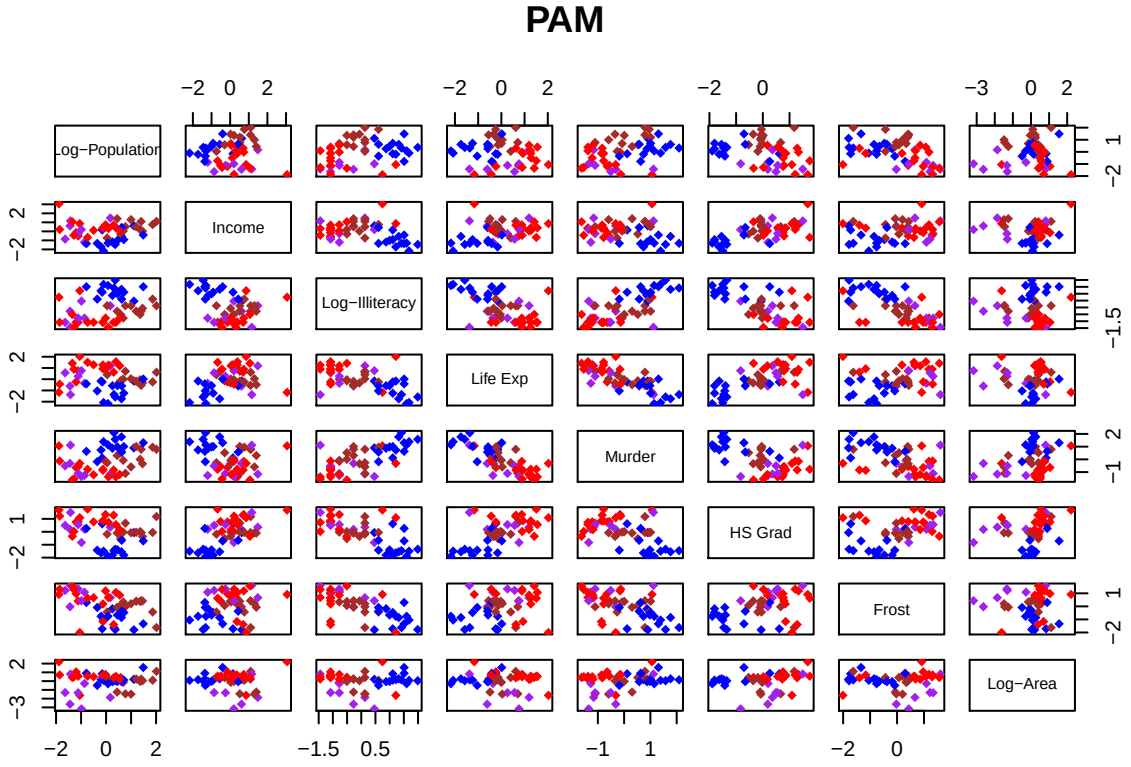
```
cl.pam<-pam.4$clustering
cl.pam
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              2              1              1              3
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              4              4              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              2              3              3              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              1              1              4              3
```

```
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              3              2              1              3
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              4              4              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              3              1              2              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              2              3              4              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              1              1              2              4
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              2              1              2              2
```

3.- Scatter plot de la matriz con los grupos

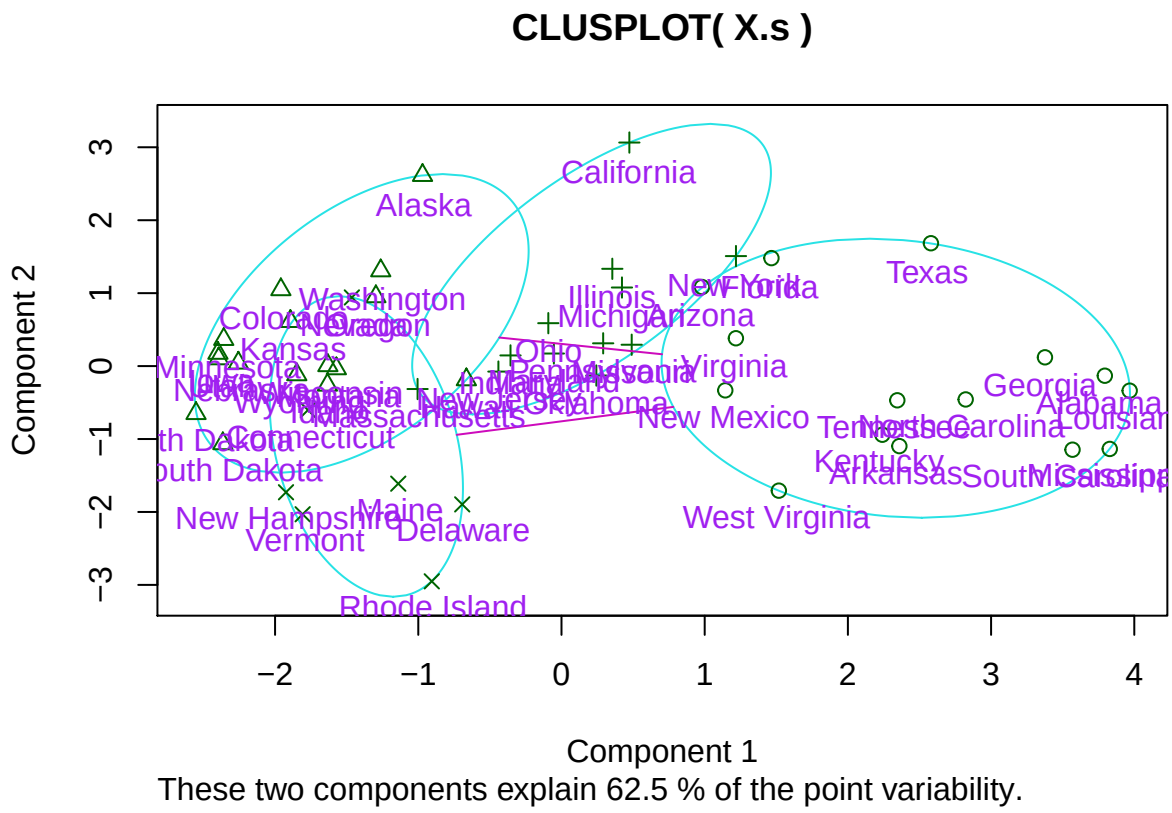
```
col.cluster<-c("blue","red", "brown", "purple")[cl.pam]
pairs(X.s, col=col.cluster, main="PAM", pch=18)
```



Visualizacion con Componentes Principales

```
library(cluster)
```

```
clusplot(X.s,cl.pam)
text(princomp(X.s)$scores[,1:2],
     labels=rownames(X.s),pos=1, col="purple")
```



Silhouette

Representación gráfica de la eficacia de clasificación de una observación dentro de un grupo:

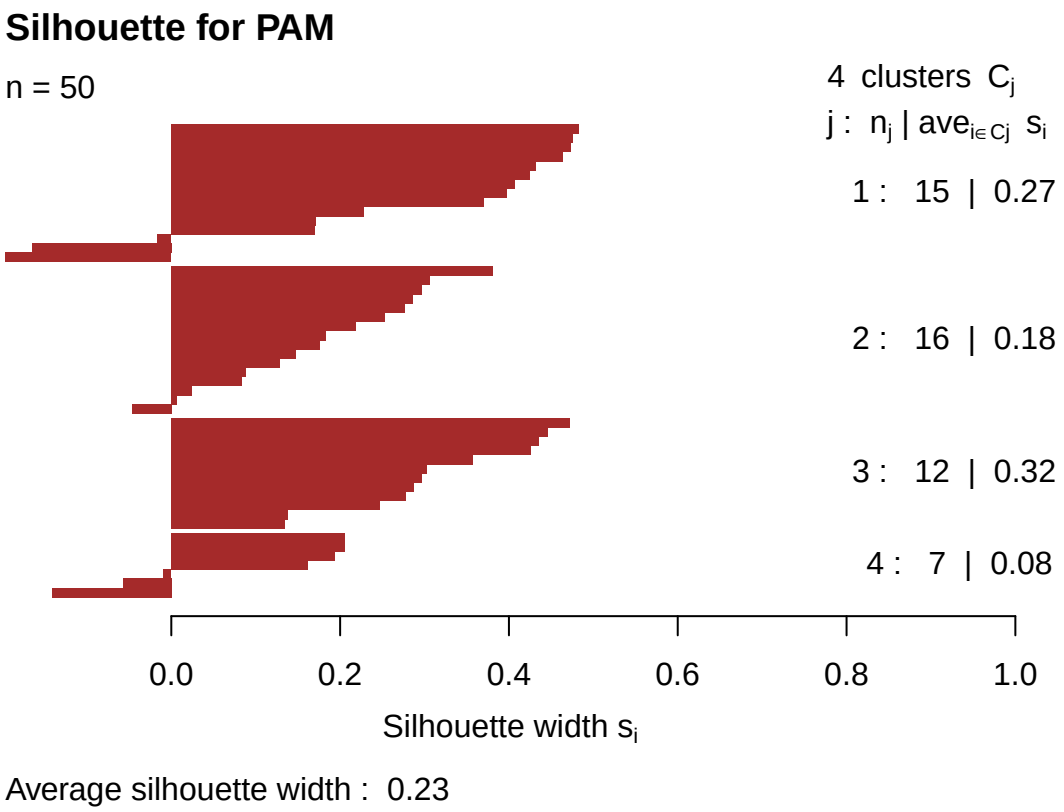
Mayor a 0.7 = mejor clasificación. Entre más cercana a 1 es mejor.

1.- Generación de los cálculos

```
dist.Euc<-dist(X.s, method = "euclidean")
Sil.pam<-silhouette(cl.pam, dist.Euc)
```

2.- Generación del gráfico

```
plot(Sil.pam, main="Silhouette for PAM",
     col="brown")
```



Con 4 clústers, hay muchos datos no clasificados. El ancho del Silhouette es de 0.23 y hay dos clústers con una probabilidad del Silhouette algo buena (0.27 y 0.32) y dos clústers con probabilidad baja (0.18 y 0.08).

Comparando el valor de cada Silhouette, el que mejor favorece al agrupamiento, son dos clúster. Su valor de Silhouette es un poco más alto: 0.26, mientras que de 3 clúster es de 0.22 y de 4 clúster es de 0.23. Todavía hay datos mal clasificados, pero si se aumenta la cantidad de clúster, puede que se dé un mejor agrupamiento, pero se corre el riesgo que el Average Silhouette sea mucho más bajo.