K-medias

Karina Itzel Rodríguez Conde

2022-05-26

K - MEDIAS

Introducción

K-medias es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en **k** grupos, en el que cada observación pertenece al grupo cuyo valor medio es más cercano

El algoritmo de k-medias busca la partición óptima con la restricción de que en cada iteración sólo se permite un elemento de un grupo a otro. En su aplicación habitual, se debe fijar el número de grupos (k) en el algoritmo.

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 k-means

1.- Separación de filas y columnas.

```
dim(X)
```

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

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

2.- Estandarización univariante.

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

3.- Algoritmo k-medias (3 grupos)

nstart: cantidad de subconjuntos aleatorios que se escogen para realizar los cálculos de algoritmo.

```
Kmeans.3<-kmeans(X.s, 3, nstart=25)
```

centroides:

```
Kmeans.3$centers
```

```
##   Log-Population     Income Log-Illiteracy   Life Exp      Murder    HS Grad
## 1      0.5693805  0.5486843     0.05412021  0.1388564 -0.01977495  0.1203417
## 2      0.2360549 -1.2266128     1.31921387 -1.0778757  1.10983501 -1.3566922
## 3     -0.7900149  0.2080926    -0.93960948  0.5642988 -0.71791785  0.7707484
##         Frost   Log-Area
## 1 -0.3291597 -0.4878988
## 2 -0.7719510  0.1991243
## 3  0.8803670  0.4093602
```

clúster de pertenencia:

```
Kmeans.3$cluster
```

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

4.- SCDG

```
SCDG<-sum(Kmeans.3$withinss)
SCDG
```

```
## [1] 203.2068
```

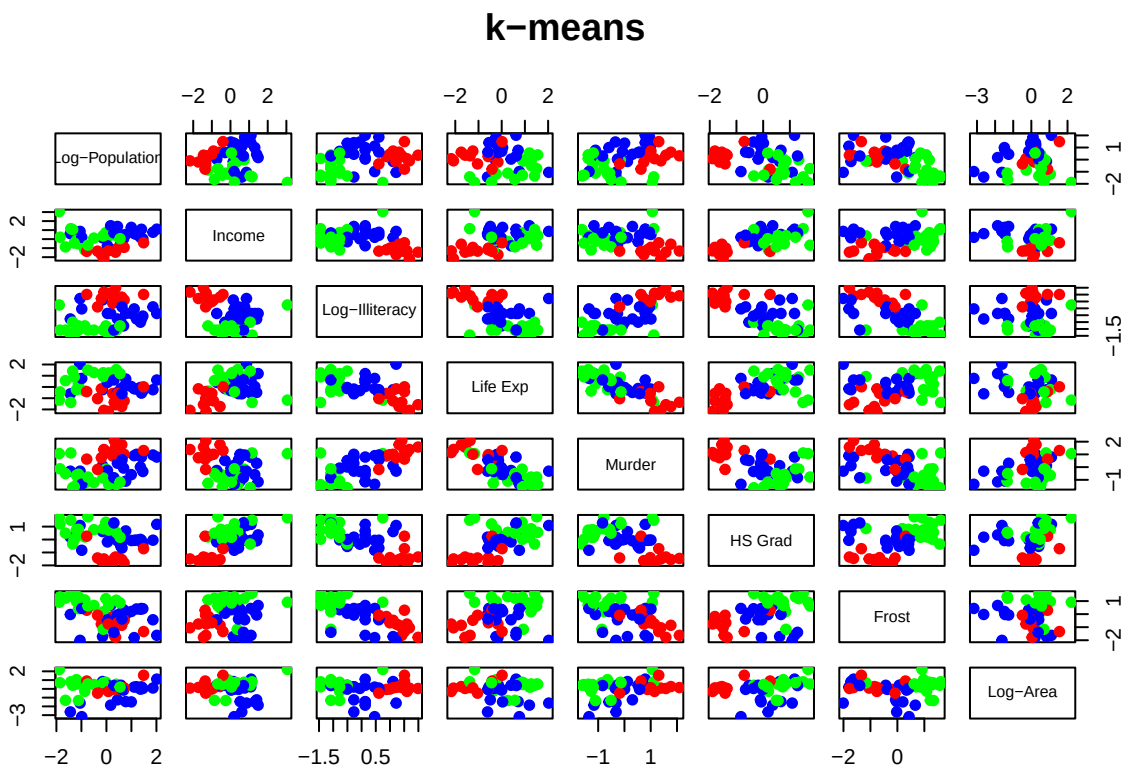
5.- Clústers

```
cl.kmeans<-Kmeans.3$cluster
cl.kmeans
```

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

6.- Scatter plot con la división de grupos obtenidos (se utiliza la matriz de datos centrados).

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

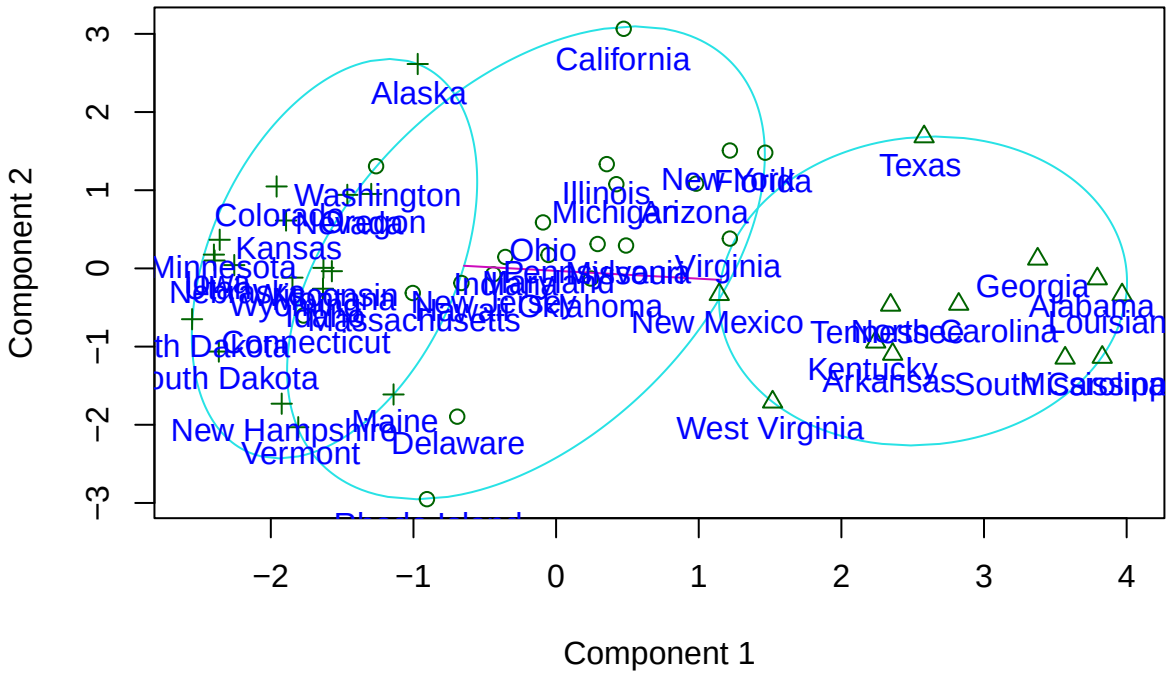


Visualizacion con los dos componentes principales

```
library(cluster)
```

```
clusplot(X.s, cl.kmeans,
         main="Dos primeras componentes principales")

text(princomp(X.s)$score[,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:

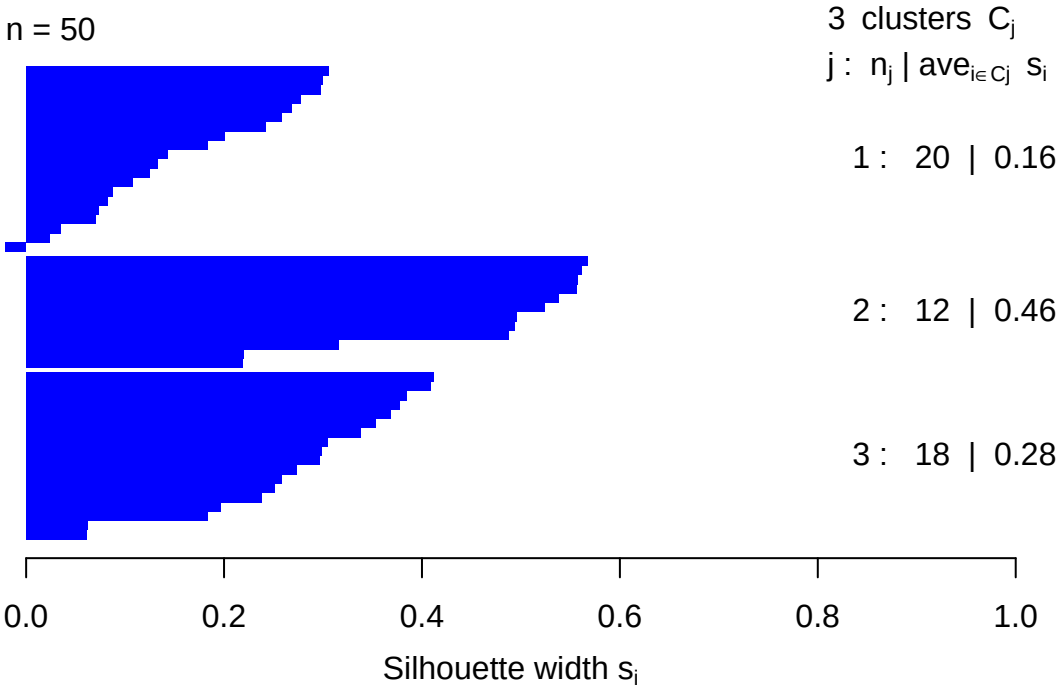
Mayor a 0.7 = mejor clasificación. Entre más cercana a 1 es mejor.

1.- Generación de los cálculos

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

2.- Generación del gráfico

```
plot(Sil.kmeans, main="Silhouette for k-means",
col="blue")
```



Dado el gráfico anterior: el clúster 1 contiene 20 estados y una probabilidad de Silhouette del 0.16, considerada como baja. El clúster 2 contiene 12 estados y su probabilidad de Silhouette es del 0.46, comparado con el clúster 1, su probabilidad es buena. Mientras que, el clúster 3 contiene 18 estados y su probabilidad es del 0.28, considerado como bajo. Hay un dato que no se clasifica, es negativo y como el valor del Silhouette es de 0.28, 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.- Algoritmo k-medias (2 grupos)

nstart: cantidad de subconjuntos aleatorios que se escogen para realizar los cálculos de algoritmo.

```
Kmeans.2<-kmeans(X.s, 2, nstart=25)
```

centroides:

```
Kmeans.2$centers
```

```
##   Log-Population     Income Log-Illiteracy   Life Exp     Murder    HS Grad
## 1      0.3921592 -0.7973132      1.1635825 -0.8863645  0.9913208 -1.0270524
## 2     -0.1845455  0.3752062     -0.5475682  0.4171127 -0.4665039  0.4833188
##        Frost   Log-Area
## 1 -0.8493032  0.2164565
## 2  0.3996721 -0.1018619
```

clúster de pertenencia:

```
Kmeans.2$cluster
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
```

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

2.- SCDG

```
SCDG<-sum(Kmeans.2$withinss)
SCDG
```

```
## [1] 257.0639
```

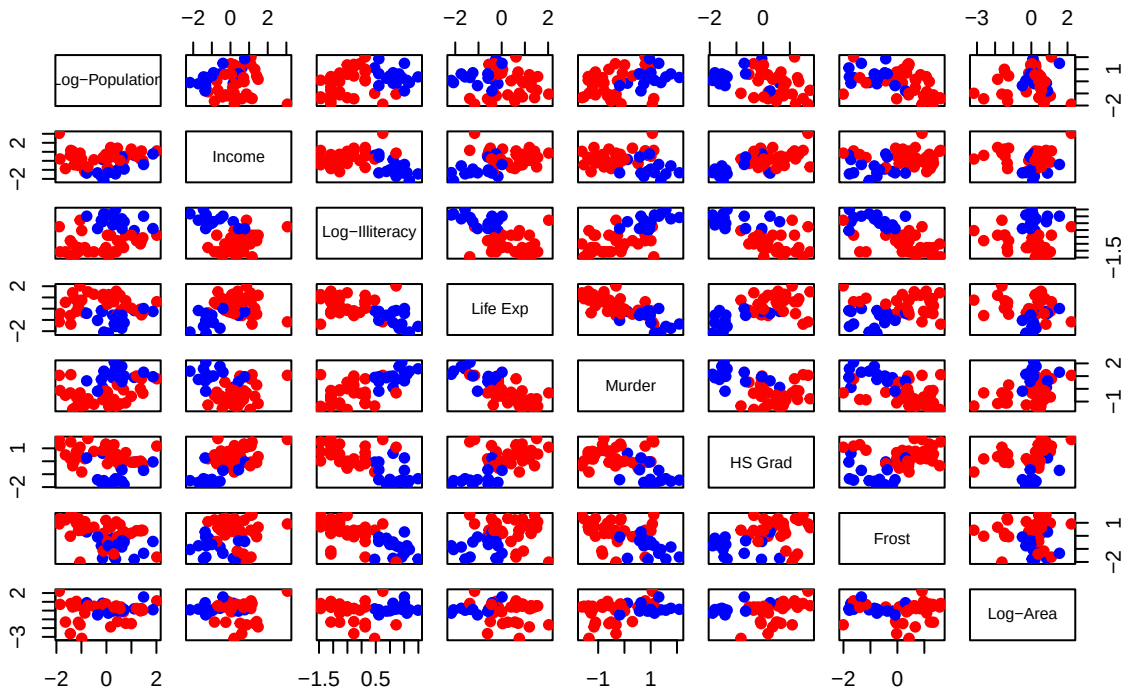
3.- Clústers

```
cl.kmeans<-Kmeans.2$cluster
cl.kmeans
```

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

4.- Scatter plot con la división de grupos obtenidos (se utiliza la matriz de datos centrados).

```
col.cluster<-c("blue", "red")[cl.kmeans]
pairs(X.s, col=col.cluster, main="k-means", pch=19)
```

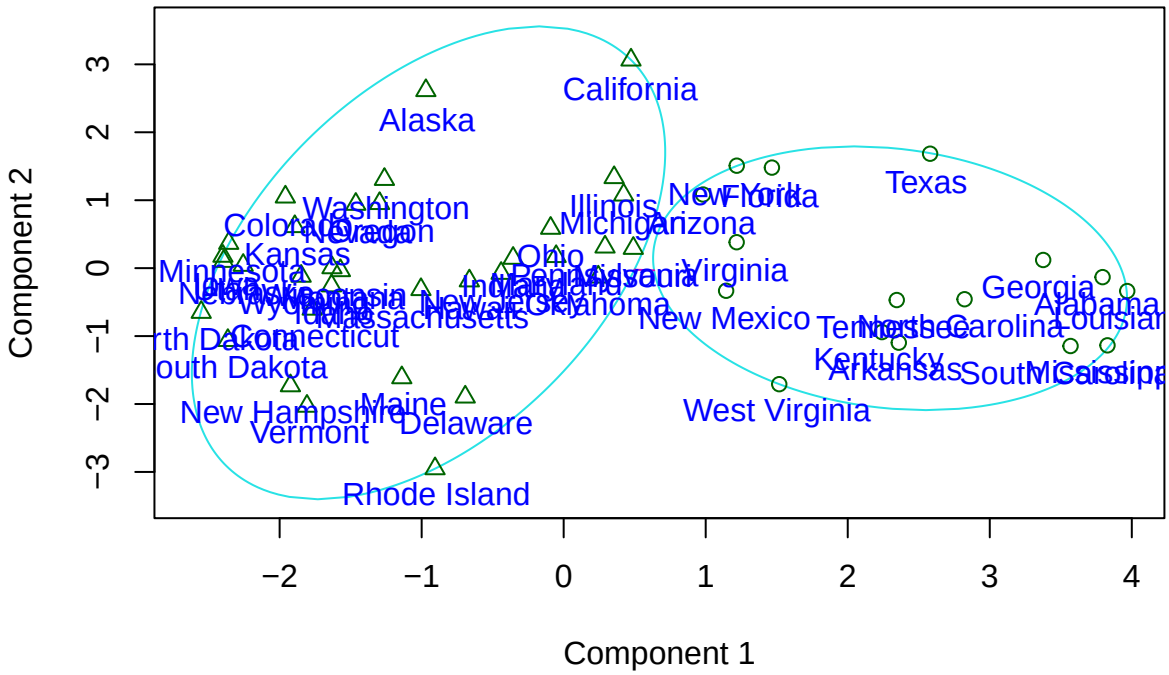


Visualizacion con los dos componentes principales

```
library(cluster)
```

```
clusplot(X.s, cl.kmeans,
         main="Dos primeras componentes principales")

text(princomp(X.s)$score[,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:

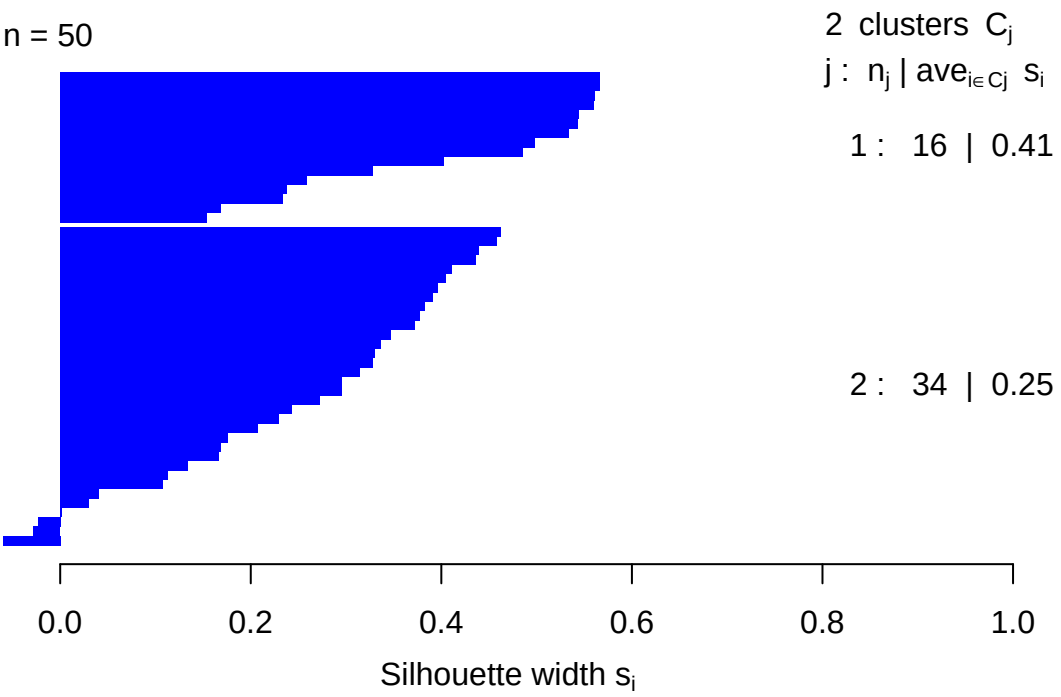
Mayor a 0.7 = mejor clasificación. Entre más cercana a 1 es mejor.

1.- Generación de los cálculos

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

2.- Generación del gráfico

```
plot(Sil.kmeans, main="Silhouette for k-means",
col="blue")
```



Average silhouette width : 0.3

Con base al gráfico anterior, con 2 clústers, hay 3 datos no clasificados, pero el ancho del valor del Silhouette es de 0.3. Hay un clúster con una probabilidad del Silhouette buena y otra con probabilidad un poco baja pero aceptable.

Se probará ahora con 4 clúster.

4 Clústers

1.- Algoritmo k-medias (4 grupos)

nstart: cantidad de subconjuntos aleatorios que se escogen para realizar los cálculos de algoritmo.

```
Kmeans.4<-kmeans(X.s, 4, nstart=25)
```

centroides:

```
Kmeans.4$centers
```

```
##   Log-Population     Income Log-Illiteracy   Life Exp     Murder     HS Grad
## 1     -0.1575882  0.9109826      0.2165582  0.5182427 -0.6480455  0.18472210
## 2     -0.7325785  0.2338173     -0.9470331  0.5675879 -0.7240168  0.79789938
## 3      0.1223312 -1.3014617      1.3019262 -1.1773136  1.0919809 -1.41578257
## 4      1.0520357  0.2689748      0.1658871 -0.1124169  0.4831422 -0.06765652
##         Frost    Log-Area
## 1 -0.1187800 -1.92526117
## 2  0.7606648  0.40780454
## 3 -0.7206500  0.07602772
## 4 -0.4380016  0.37632593
```

clúster de pertenencia:

```
Kmeans.4$cluster
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              3              2              4              3              4
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              1              1              4              3
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              2              4              4              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              3              3              2              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              4              2              3              4
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              2              2              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              4              3              2              4
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              4              2              4              1              3
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              3              4              2              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              4              2              3              2              2
```

2.- SCDG

```
SCDG<-sum(Kmeans.4$withinss)
SCDG
```

```
## [1] 167.0685
```

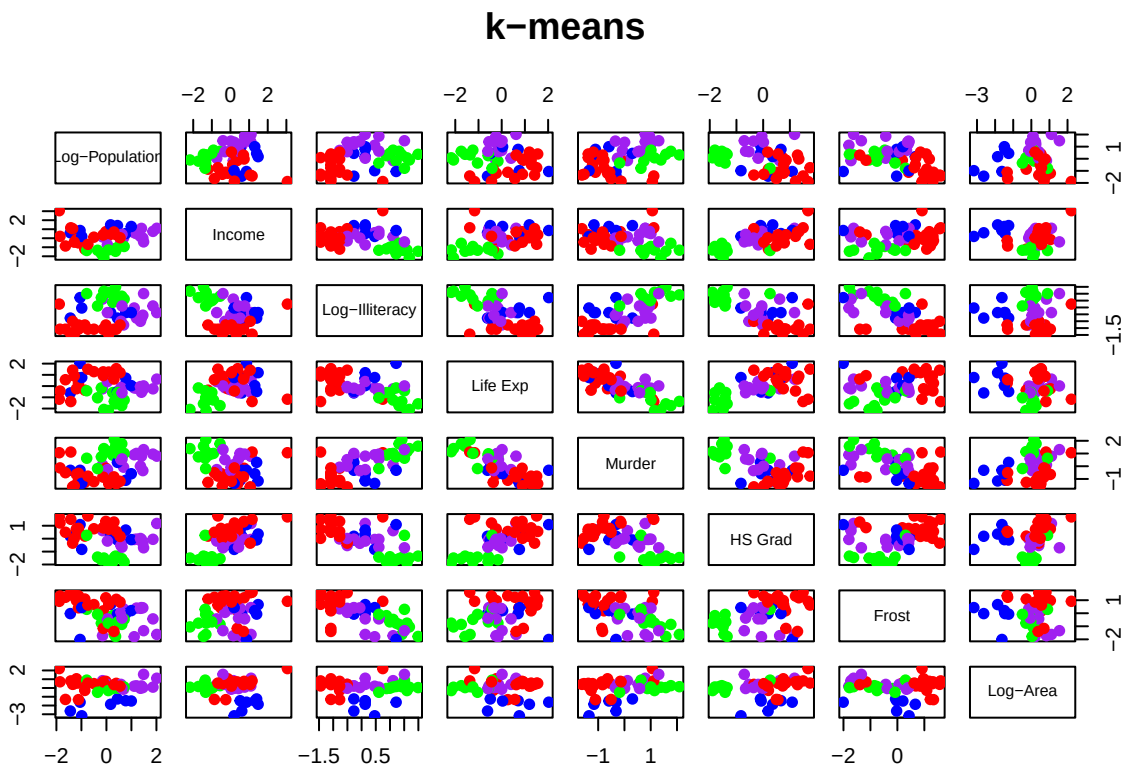
3.- Clústers

```
cl.kmeans<-Kmeans.4$cluster
cl.kmeans
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              3              2              4              3              4
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              1              1              4              3
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              2              4              4              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              3              3              2              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              4              2              3              4
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              2              2              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              4              3              2              4
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              4              2              4              1              3
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              3              4              2              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              4              2              3              2              2
```

4.- Scatter plot con la división de grupos obtenidos (se utiliza la matriz de datos centrados).

```
col.cluster<-c("blue", "red", "green", "purple")[cl.kmeans]
pairs(X.s, col=col.cluster, main="k-means", pch=19)
```

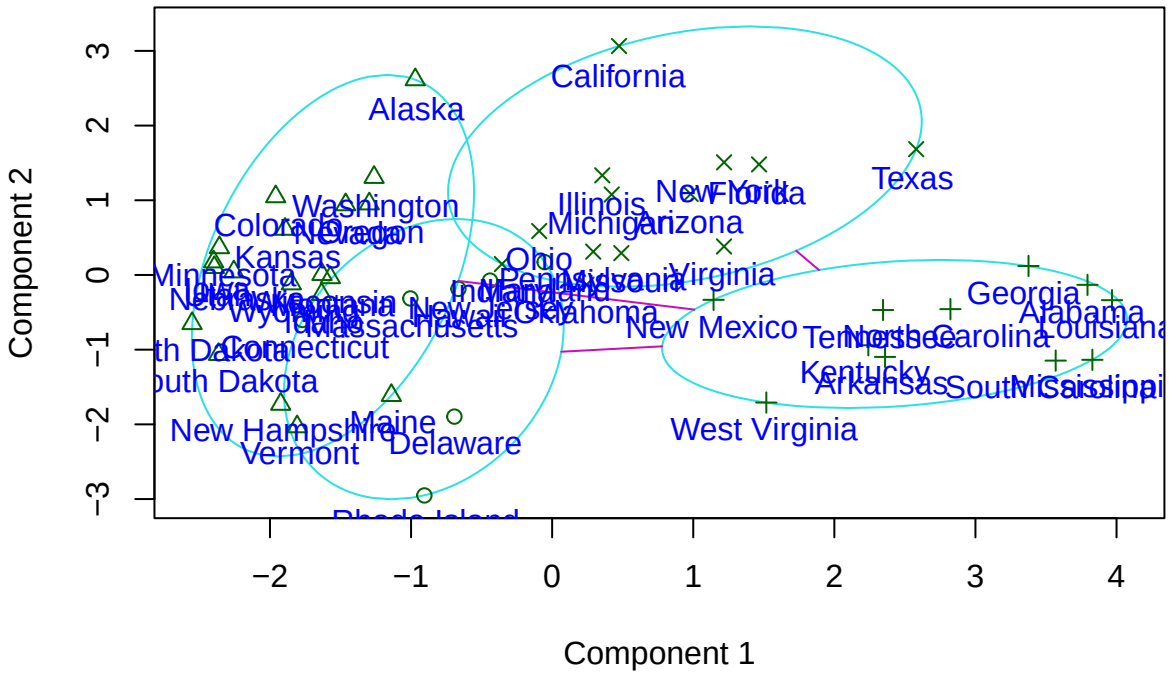


Visualizacion con los dos componentes principales

```
library(cluster)
```

```
clusplot(X.s, cl.kmeans,
         main="Dos primeras componentes principales")

text(princomp(X.s)$score[,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:

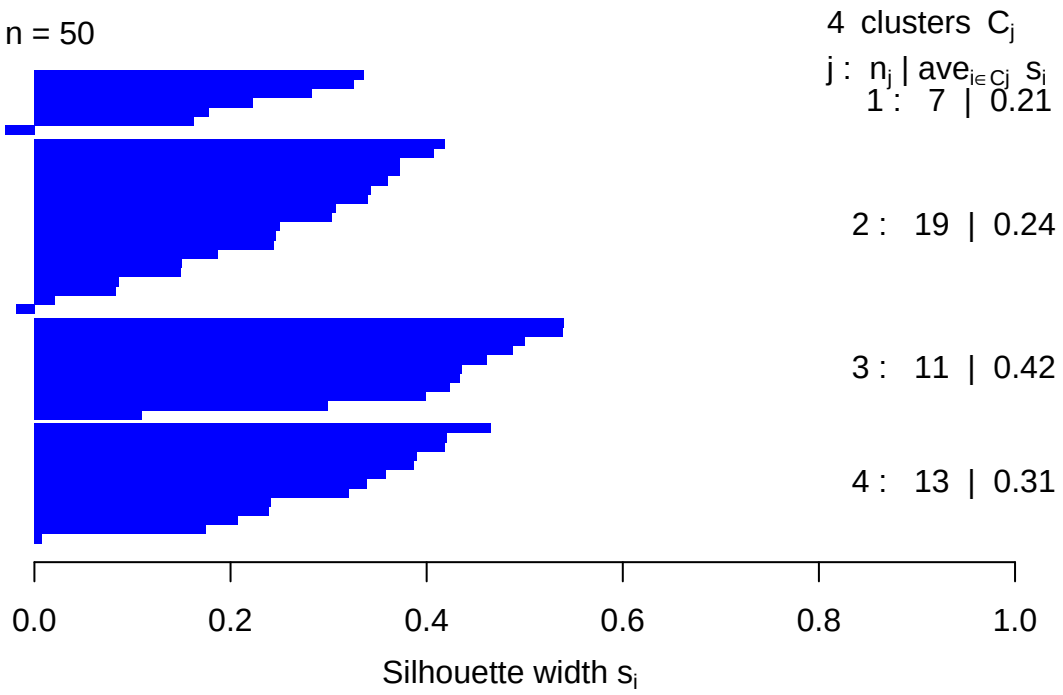
Mayor a 0.7 = mejor clasificación. Entre más cercana a 1 es mejor.

1.- Generación de los cálculos

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

2.- Generación del gráfico

```
plot(Sil.kmeans, main="Silhouette for k-means",
col="blue")
```



Average silhouette width : 0.29

Con 4 clústers, hay 2 datos no clasificados. El ancho del Silhouette es de 0.29 y hay dos clústers con una probabilidad del Silhouette buena y dos clústers con probabilidad un poco baja.

Por ello, la mejor clasificación, la que favorece por el valor del ancho del Silhouette y por sus valores de probabilidad del Silhouette, sería un clúster de 2 grupos. Esto se debe a que, el ancho del Silhouette es de 0.3 y es mayor que del clúster 3 y 4, además de que, la probabilidad de los dos grupos es buena.