K-MEDIAS

Arleth Michell Morales García

2022-05-26

Cargar la matriz de datos

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

Nombre de las variables

```
colnames(X)
```

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

TRANSFORMACIÓN DE DATOS

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-MEDIAS

1. Separación de filas y columnas

```
dim(X)
```

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

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

2. Estandarización univariante

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

3. Algoritmo k-medias (*3 grupos*)

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

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

Visualizar centroides

```r
Kmeans.3$centers
```

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

Cluster de pertenencia

```r
Kmeans.3$cluster
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              3              2              1              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              3              2              2              2              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              3              2              2              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              1              1              3              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              2              3              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              3              3              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
```

```
##              1              2              1              3              2
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              3              2              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              1              1              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              1              3              3
```

Aquí se puede observar a qué cluster pertenece cada ciudad.

4. SCDG

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

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

Lo mínimo que puede minimizar el algoritmo es 203.20

5. CLUSTERS

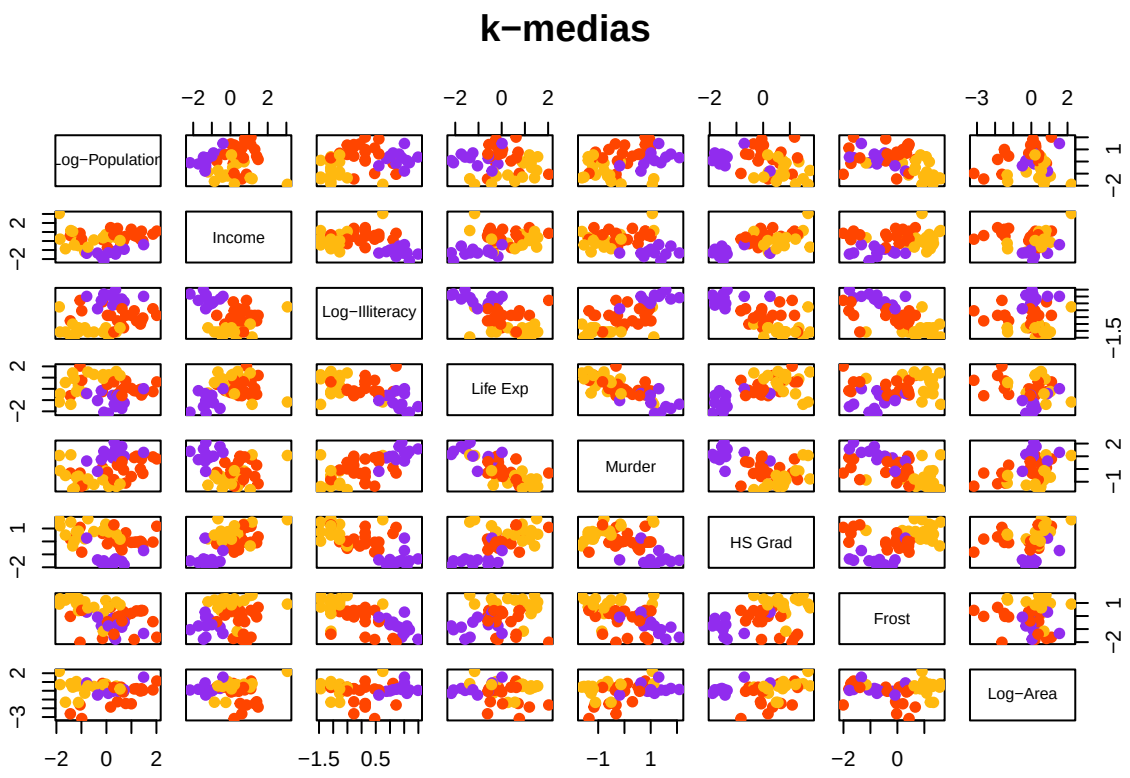
```
cl.kmeans<-Kmeans.3$cluster
cl.kmeans
```

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

6. Scatter plot con la division de grupos obtenidos

(se utiliza la matriz de datos centrados).

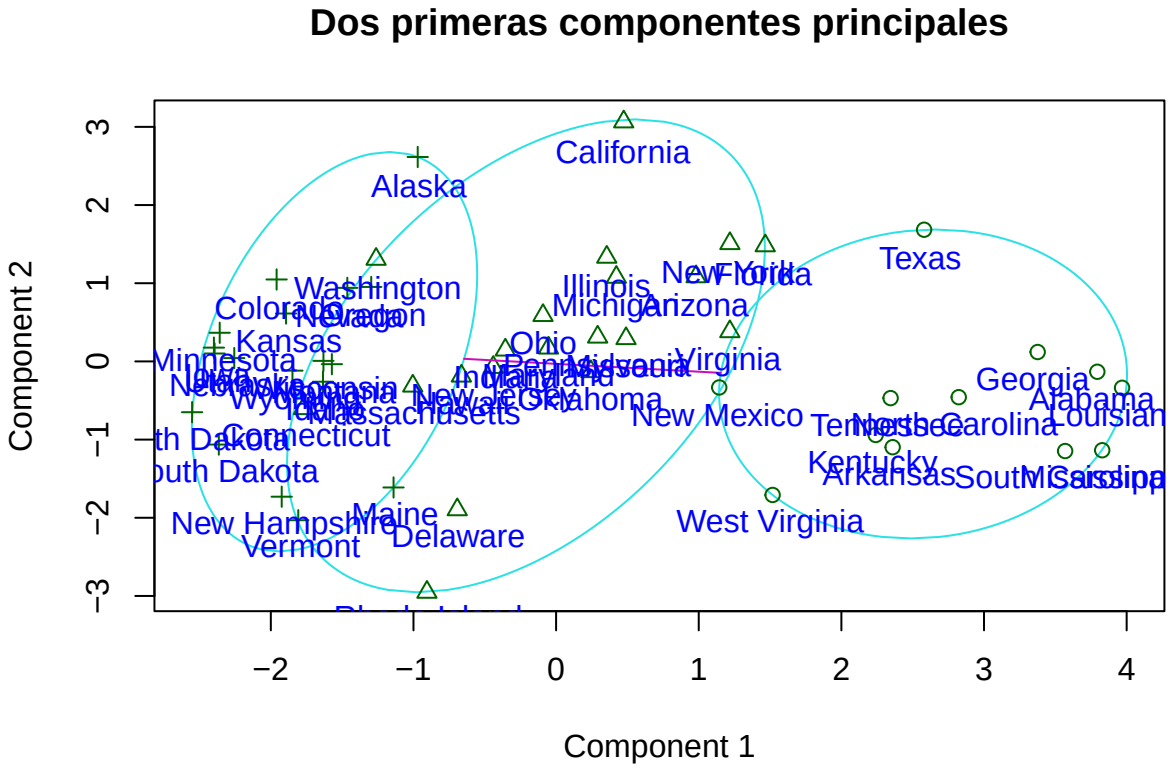
```
col.cluster<-c("purple2", "orangered", "darkgoldenrod1")[cl.kmeans]
pairs(X.s, col=col.cluster, main="k-medias", pch=19)
```



VISUALIZACIÓN CON LAS 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.

Podemos observar que Alaska pertenece al cluster 3, California cluster 1 y Texas cluster 2

SILHOUETTE

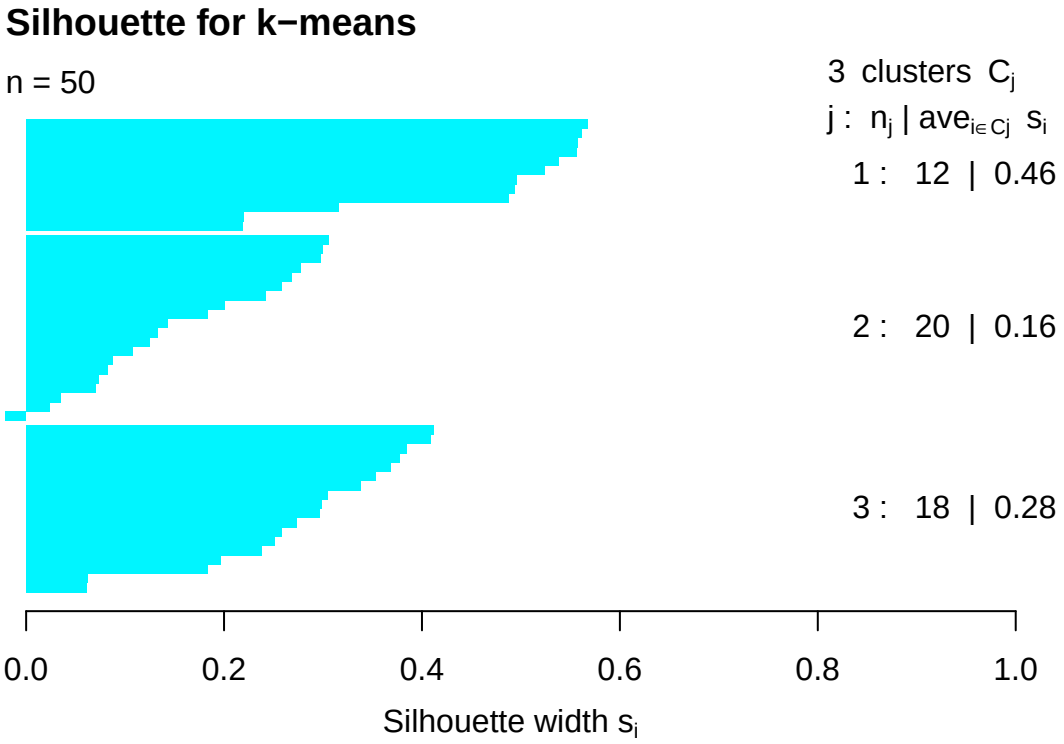
Representación gráfica de la eficacia de clasificación de una observación dentro de un grupo.

1. Generación de los calculos

```
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="turquoise1")
```



Silhouette va de 0 a 1, si está en 0 no está clasificando bien

En el grupo uno está muy bajita la clasificación con 0.16.

El cluster 2 "presenta" mejor clasificación.

Con estos resultados nos demuestra que no fue tan bueno

tomar 3 clusters.

Cargar la matriz de datos

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

Nombre de las variables

```
colnames(X)
```

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

TRANSFORMACIÓN DE DATOS

1. Transformación de las variables $x1, x3yx8$ 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-MEDIAS

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 (*5 grupos*)

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

```
Kmeans.5<-kmeans(X.s, 5, nstart=30)
```

Se realizarán 5 clusters.

Visualizar centroides

```
Kmeans.5$centers
```

```
##   Log-Population        Income Log-Illiteracy   Life Exp     Murder     HS Grad
## 1     -0.5470524  0.0007323385     -1.0134235  0.8605152 -0.9878669  0.67299139
## 2     -0.1575882  0.9109826094      0.2165582  0.5182427 -0.6480455  0.18472210
## 3      0.1223312 -1.3014616989      1.3019262 -1.1773136  1.0919809 -1.41578257
## 4      1.0520357  0.2689747904      0.1658871 -0.1124169  0.4831422 -0.06765652
## 5     -1.7220507  1.4769369102     -0.5929507 -0.9946909  0.6831838  1.46407534
##        Frost    Log-Area
## 1  0.6632731  0.25141793
## 2 -0.1187800 -1.92526117
## 3 -0.7206500  0.07602772
## 4 -0.4380016  0.37632593
## 5  1.2800868  1.24186646
```

Cluster de pertenencia

```
Kmeans.5$cluster
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              3              5              4              3              4
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              1              2              2              4              3
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              1              4              4              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              3              3              1              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              4              1              3              4
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              1              5              1              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              4              3              1              4
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              4              1              4              2              3
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              3              4              1              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              4              1              3              1              5
```

Aquí se puede observar a qué cluster pertenece cada ciudad.

4. SCDG

```
SCDG<-sum(Kmeans.5$withinss)
SCDG
```

```
## [1] 136.8587
```

Lo mínimo que puede minimizar el algoritmo es 136.8587

5. CLUSTERS

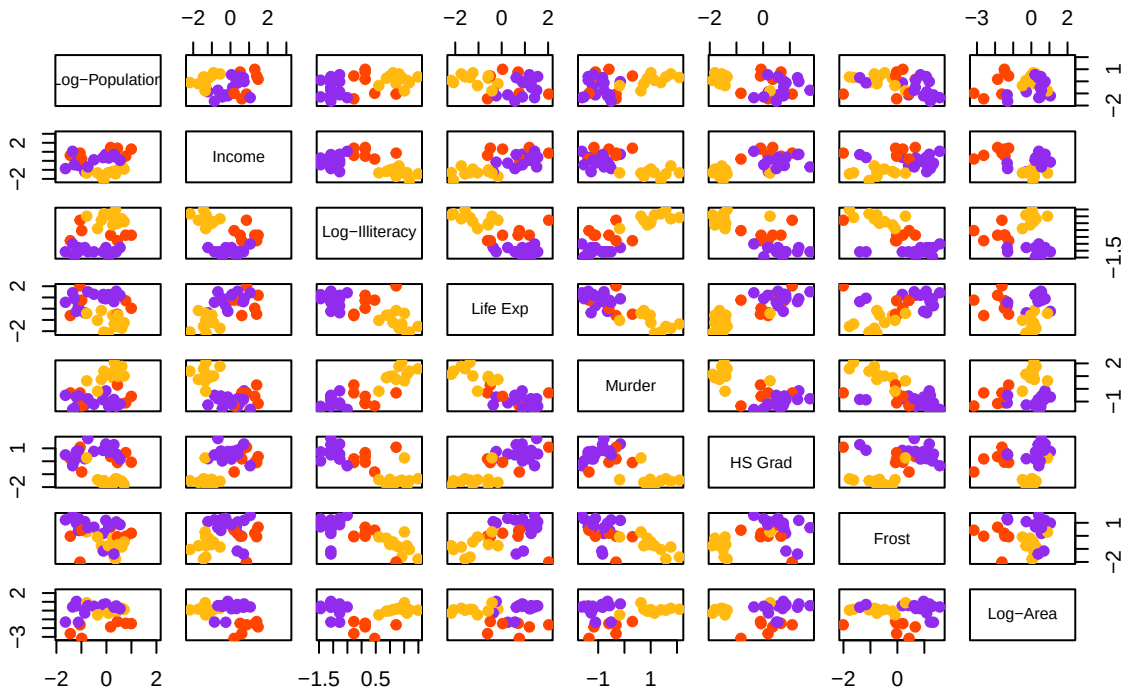
```
cl.kmeans<-Kmeans.5$cluster
cl.kmeans
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              3              5              4              3              4
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              1              2              2              4              3
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              1              4              4              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              3              3              1              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              4              1              3              4
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              1              5              1              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              4              3              1              4
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              4              1              4              2              3
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              3              4              1              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              4              1              3              1              5
```

6. Scatter plot con la division de grupos obtenidos

(se utiliza la matriz de datos centrados).

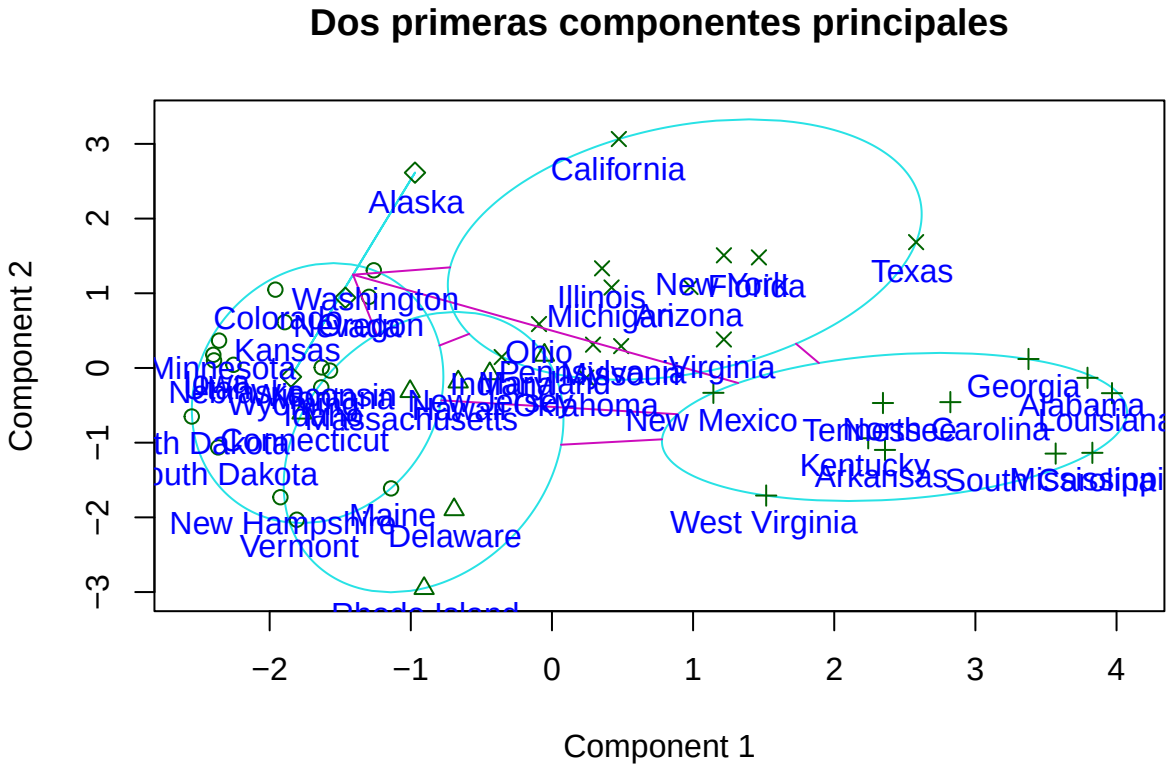
```
col.cluster<-c("purple2", "orangered", "darkgoldenrod1")[cl.kmeans]
pairs(X.s, col=col.cluster, main="k-medias", pch=19)
```



VISUALIZACIÓN CON LAS 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.

Se observa que Alaska, que pertenece al cluster 5 se sale de este agrupamiento, el cluster 5 no se visualiza bien agrupado

SILHOUETTE

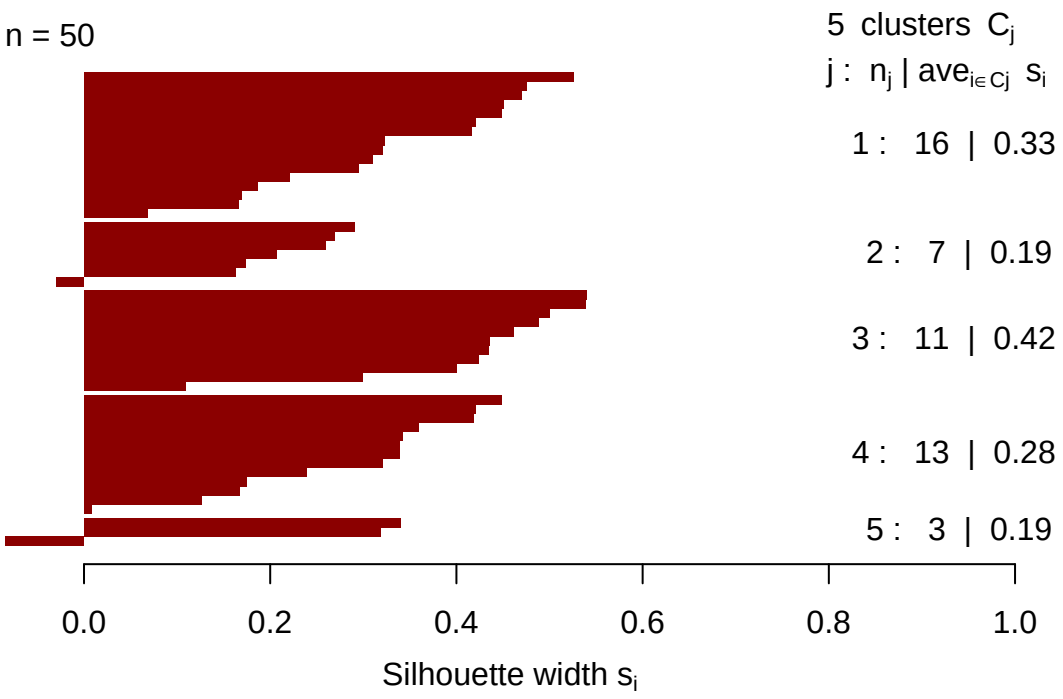
Representación gráfica de la eficacia de clasificación de una observación dentro de un grupo.

1. Generación de los calculos

```
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="red4")
```



Average silhouette width : 0.31

En el gráfico Silhouette se observa que el cluster 3 tiene una clasificación moderada en comparación con los demás cluster.

El cluster 1, 2, y 5 tienen una clasificación muy baja.

Cómo conclusión, no fue una buena idea tomar 5 clusters, ya que se dispersan muchoy no se clasifican bien, hasta se salen y la clasificación es muy baja.

Es mejor tener menos clusters.