k-means

sergio

2022-05-29

K-MEANS

Cargar la matriz de datos.

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

Transformacion de datos

1.- Transformacion de las variables x1,x3 y x8, con la funcion 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"
```

Metodo k-means

1.- Separacion de filas y columnas.

```
dim(X)
```

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

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

2.- Estandarizacion univariante.

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

3.- Algoritmo k-medias (3 grupos) cantidad de subconjuntos aleatorios que se escogen para realizar los calculos de algoritmo.

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

centroides

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

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

cluster de pertenencia

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

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

4.- SCDG

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

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

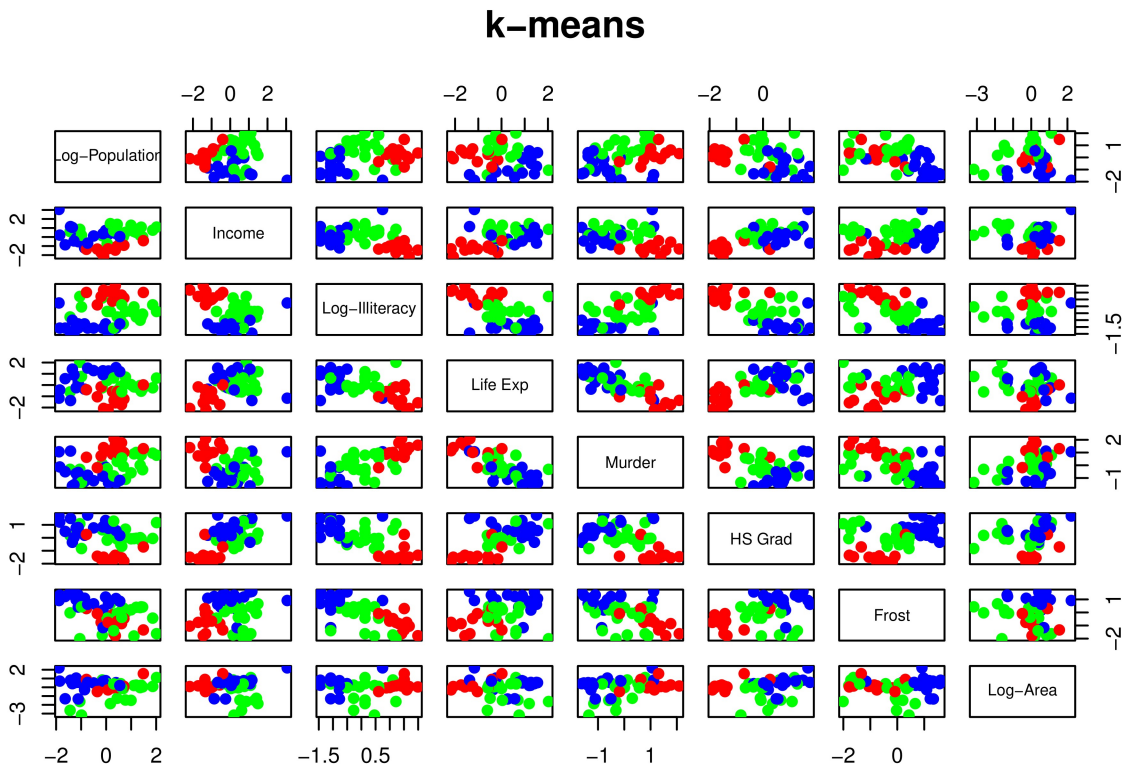
5.- Clusters

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

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

6.- Scatter plot con la division 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)
```

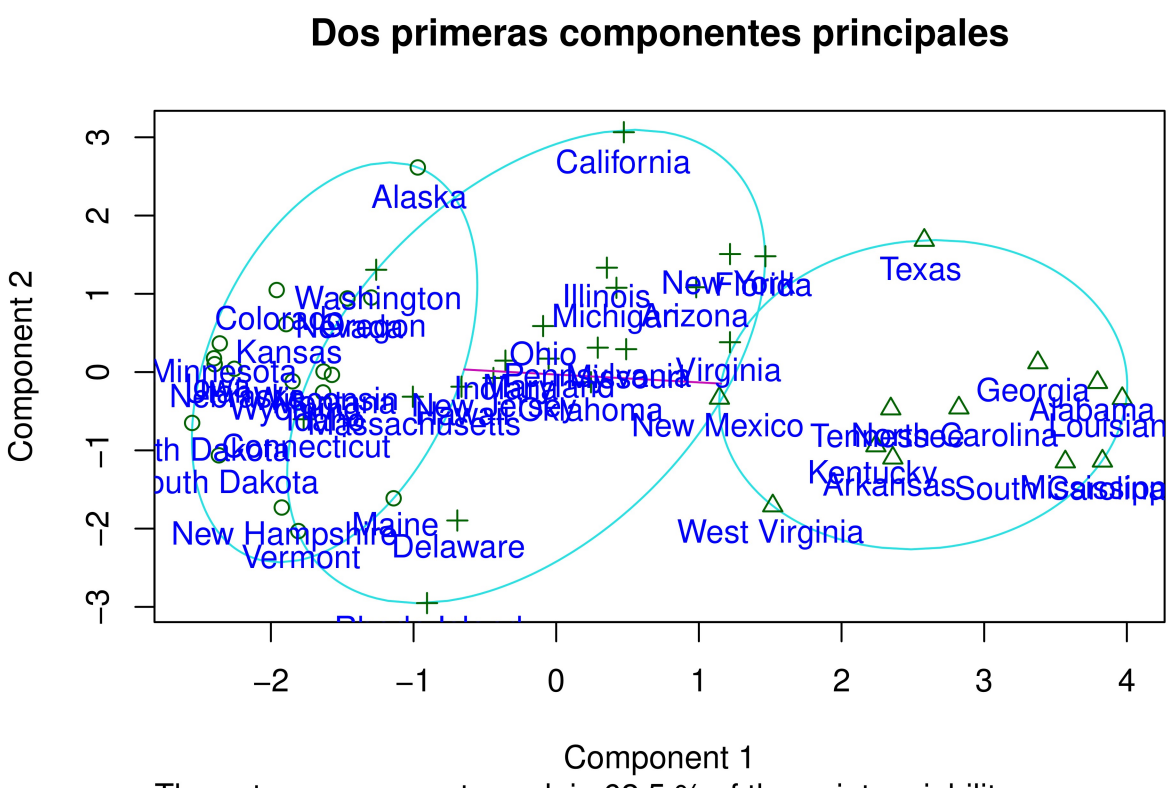


Visualizacion con las dos componentes principales

activamos librerias

```
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")
```



podemos ver como se comportan los datos y donde estan cada uno de los estados

Silhouette

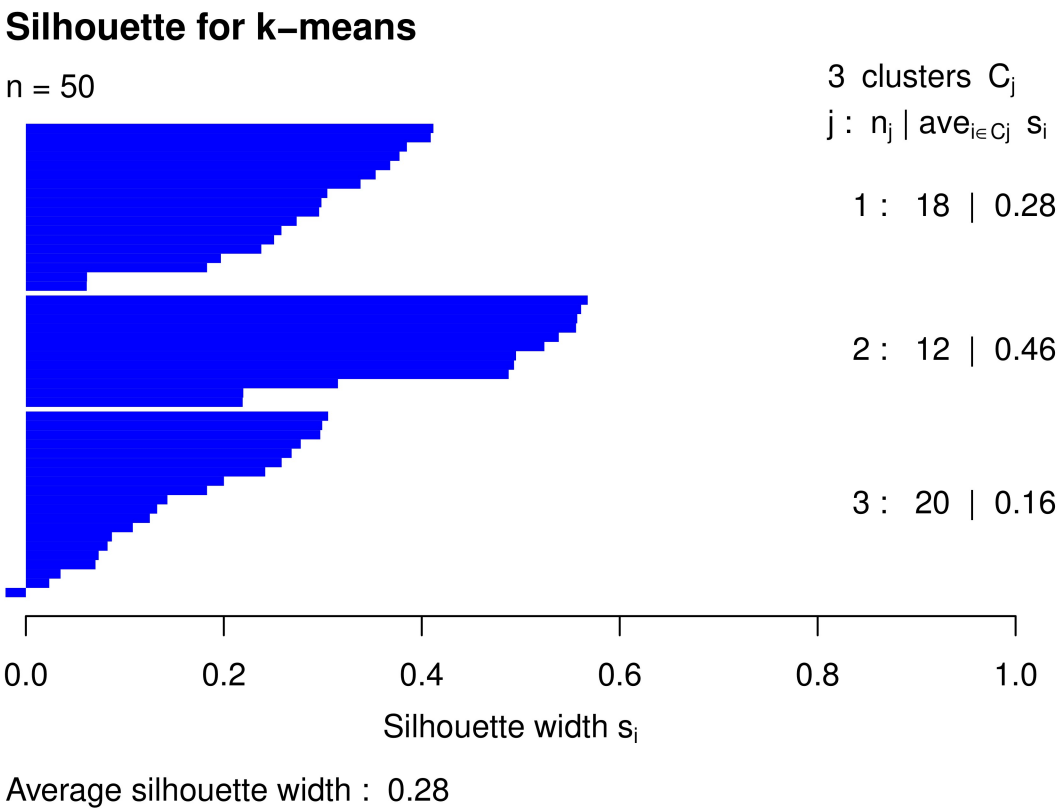
Representacion grafica de la eficacia de clasificacion de una observacion dentro de un grupo.

1.- Generacion de los calculos

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

2.- Generacion del grafico

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



podemos ver en el grafico no estan bueno porque el cluster mas grande es de .46 y no esta alto entonces no es tan bueno

EJERCICIO

1.-REPLICAR EL SCRIP PERO VAS A SUGERIR UN NUEVO NUMERO DE CLUSTERS DIFERENTES A 3 Y 1.

2.-INTERPRETASION DEL Silhouette

usando la misma base de datos voy a ver si hay un cambio en el resultado usando un par de clusters de 5 y 4

Cargar la matriz de datos.

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

Transformacion de datos

1.- Transformacion de las variables x1,x3 y x8 con la funcion de logaritmo.

```r
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"
```

Metodo k-means

1.- Separacion de filas y columnas.

```r
dim(X)
```

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

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

2.- Estandarizacion univariante.

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

3.- Algoritmo k-medias (5 grupos) cantidad de subconjuntos aleatorios que se escogen para realizar los calculos de algoritmo.

```r
Kmeans.5<-kmeans(X.s, 5, nstart=25)
```

centroides

```r
Kmeans.5$centers
```

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

cluster de pertenencia

```
Kmeans.5$cluster
```

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

4.- SCDG

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

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

5.- Clusters

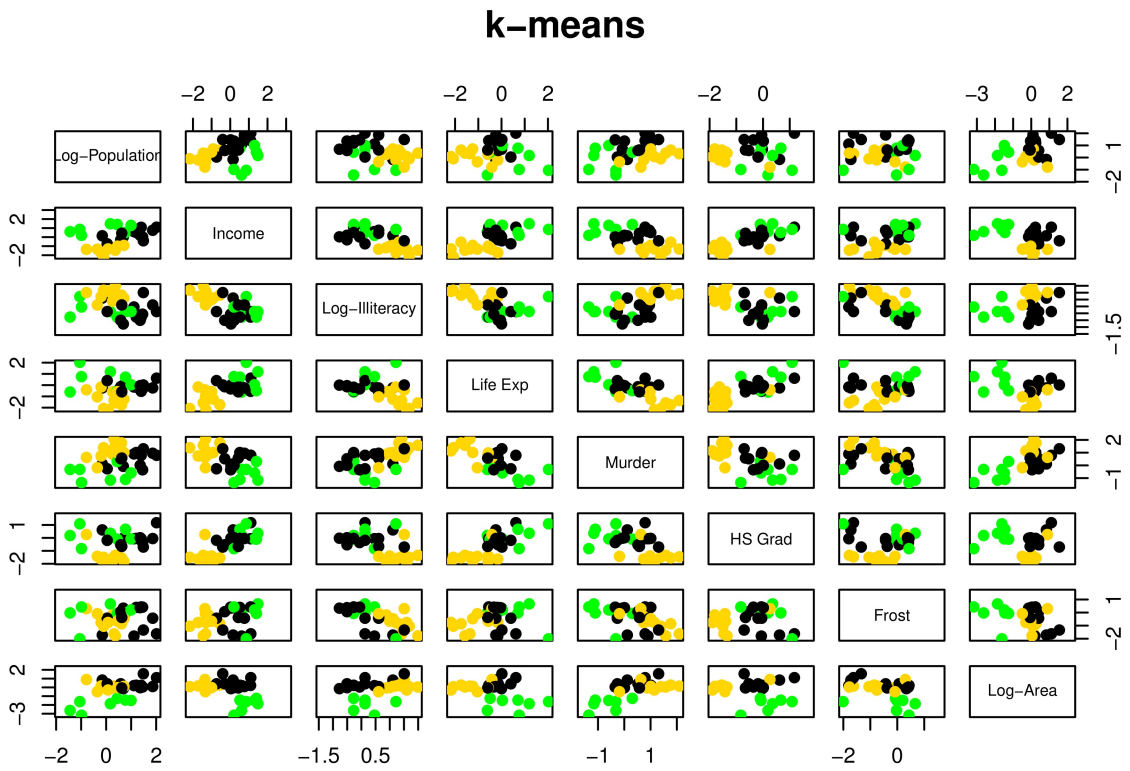
```
cl.kmeans<-Kmeans.5$cluster
cl.kmeans
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              3              4              2              3              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              5              1              1              2              3
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              5              2              2              5
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              5              3              3              5              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              2              5              3              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
```

```
##              5              5              4              5              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              2              3              5              2
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              5              2              1              3
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              5              3              2              5              5
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              5              3              5              4
```

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

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



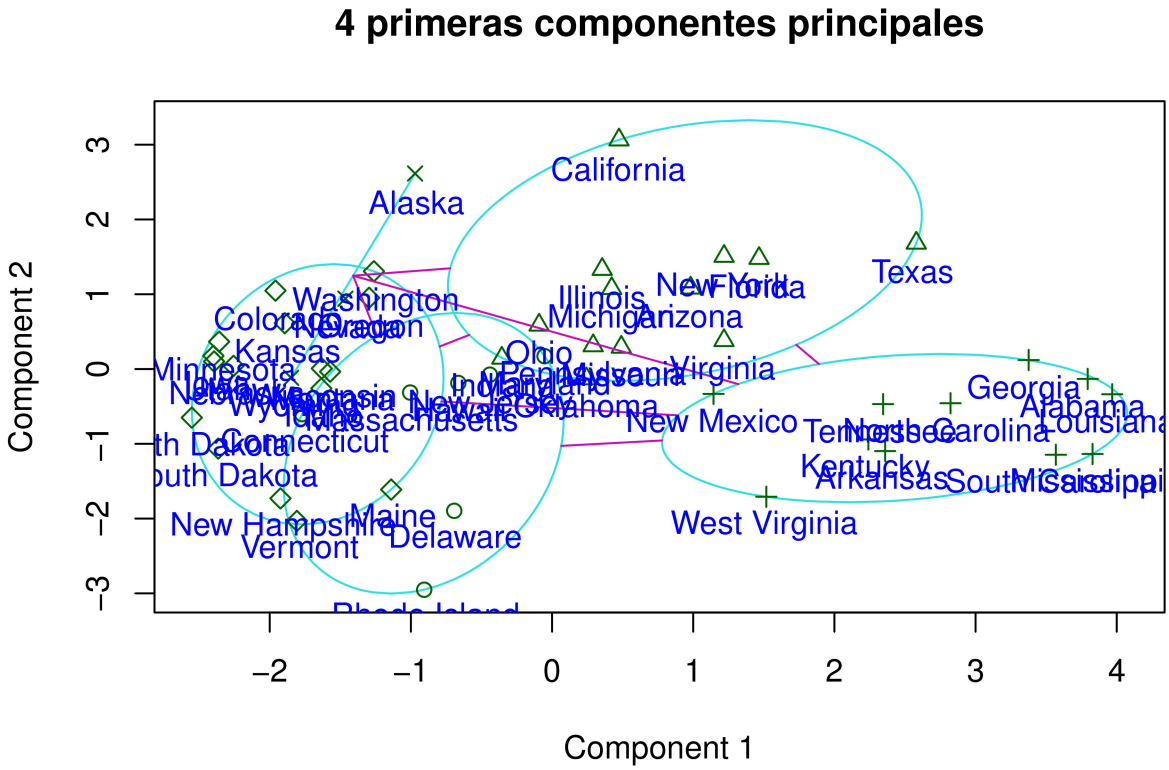
Visualizacion con las 4 componentes principales

activamos libreria y vemos los componenetes

```
library(cluster)

clusplot(X.s, cl.kmeans,
         main="4 primeras componentes principales")

text(princomp(X.s)$score[,1:4],
     labels=rownames(X.s), pos=1, col="blue")
```



These two components explain 62.5 % of the point variability.

Silhouette

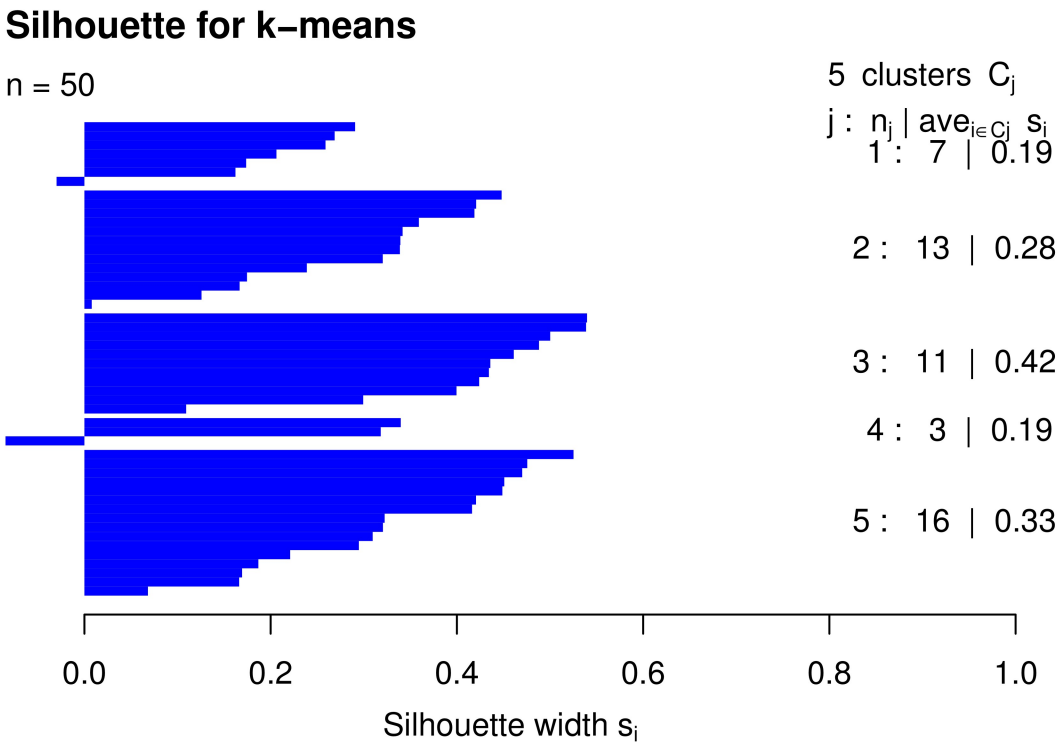
Representacion grafica de la eficacia de clasificacion de una observacion dentro de un grupo.

1.- Generacion de los calculos

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

2.- Generacion del grafico

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



bueno con el resultado de la grafica se ve mal los grupos no son mayores de 0.50 :C

Metodo k-means

1.- Separacion de filas y columnas.

```
dim(X)
```

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

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

2.- Estandarizacion univariante.

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

3.- Algoritmo k-medias (4 grupos) cantidad de subconjuntos aleatorios que se escogen para realizar los calculos de algoritmo.

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

centroides

```r
Kmeans.4$centers
```

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

cluster de pertenencia

```r
Kmeans.4$cluster
```

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

4.- SCDG

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

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

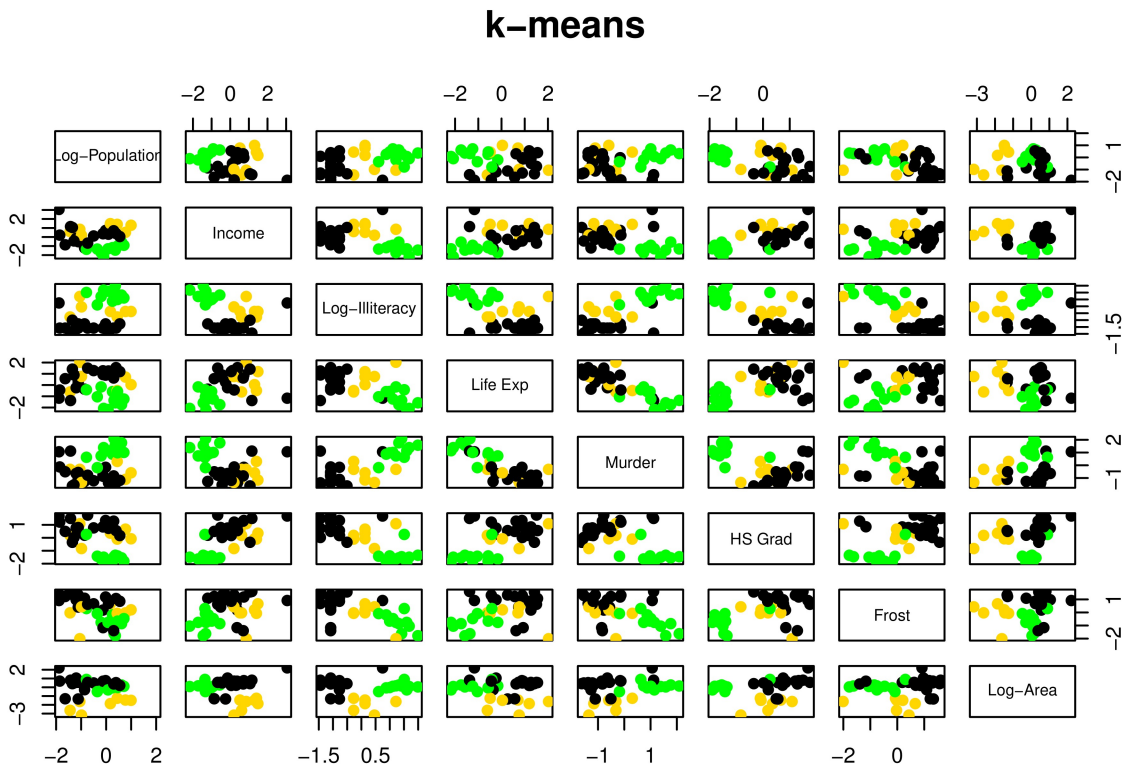
5.- Clusters

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

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

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

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



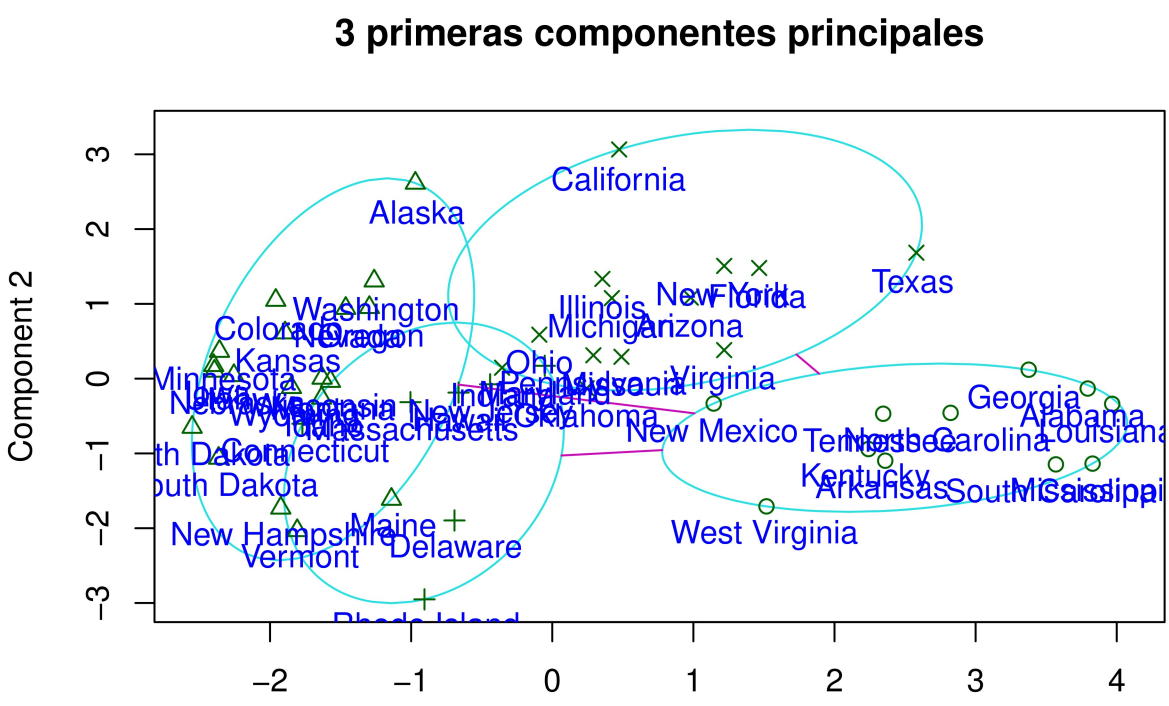
Visualizacion con las 4 componentes principales

activamos libreria y vemos los componenetes

```
library(cluster)

clusplot(X.s, cl.kmeans,
         main="3 primeras componentes principales")

text(princomp(X.s)$score[,1:3],
     labels=rownames(X.s), pos=1, col="blue")
```



These two components explain 62.5 % of the point variability.

Silhouette

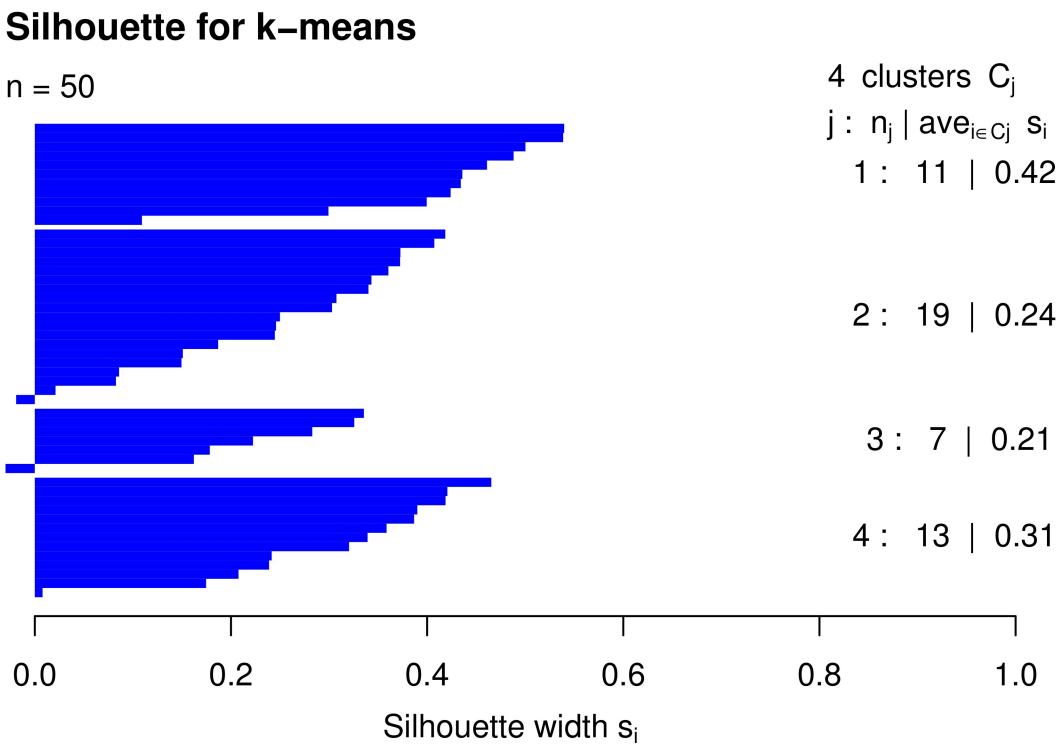
Representacion grafica de la eficacia de clasificacion de una observacion dentro de un grupo.

1.- Generacion de los calculos

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

2.- Generacion del grafico

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



bueno con el resultado de la grafica se ve mal los grupos no son mayores de 0.50 inluso es peor que si son 5 grupos :C