

K-MEANS

Ana Karen Martínez Marín

2022-05-30

Introducción

Para esta práctica se utilizó la base de datos precargada en R **state.x77**. Pero ahora se le aplicará el método K-Means.

```
## Librerías
library(cluster)
```

Matriz de datos.

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

```
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)

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
```

Cluster 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.- Clusters

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

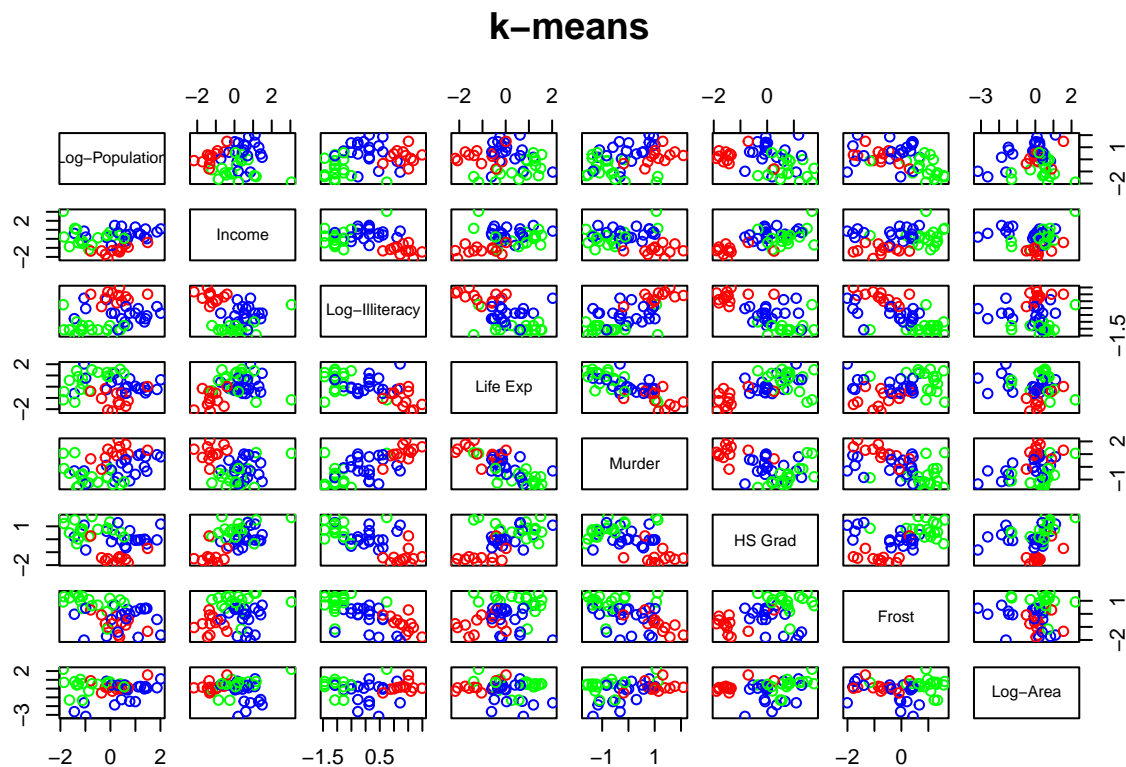
```
##      Alabama      Alaska      Arizona      Arkansas      California
##          2          3          1          2          1
##      Colorado  Connecticut      Delaware      Florida      Georgia
```

```
##           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).

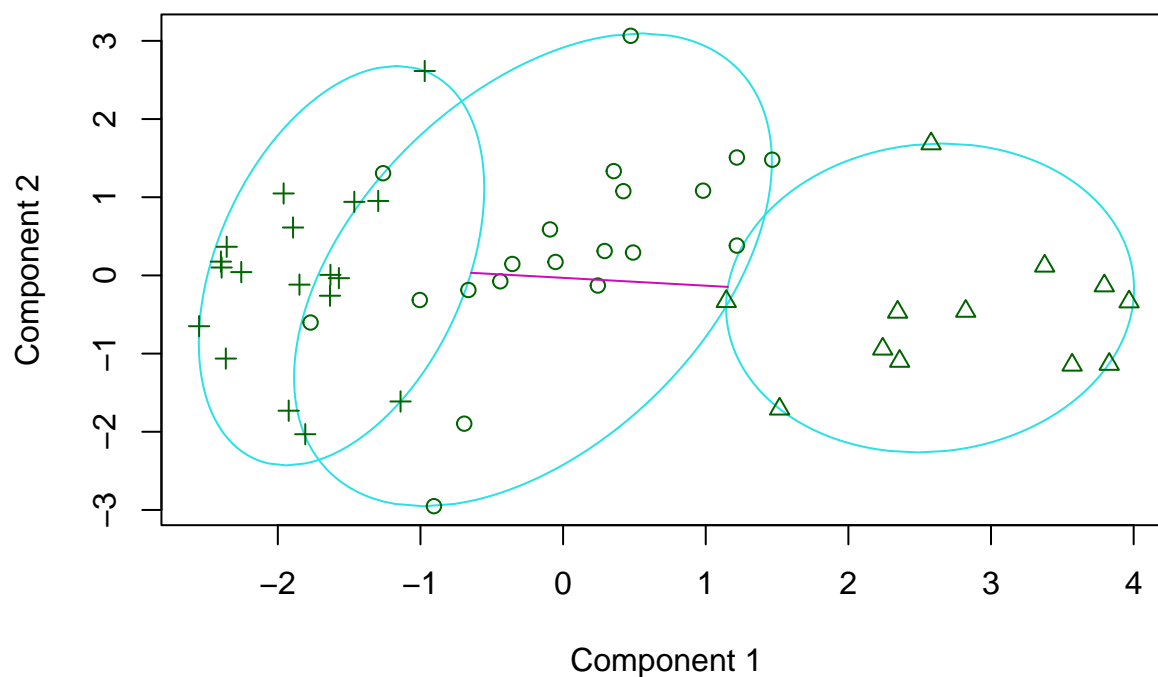
```
col.cluster<-c("blue", "red", "green")[cl.kmeans]
pairs(X.s, col=col.cluster, main="k-means", pch=21)
```



Visualización con las dos componentes principales

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

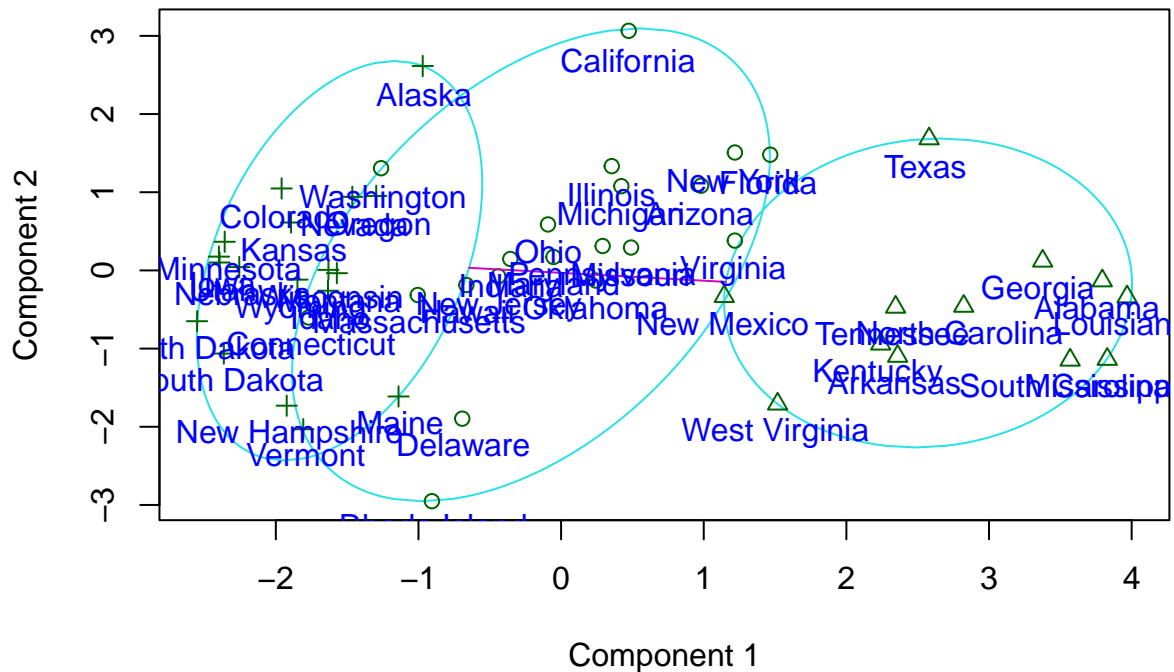
Dos primeras componentes principales



These two components explain 62.5 % of the point variability.

```
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")
```

Dos primeras componentes principales



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.- 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")
```

Silhouette for k-means

n = 50

