# K-Medias

Edgar Ortiz Mota

2022-05-28

Aquí se consieran las medianas busca k objetos representativos

```
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.

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

3.- Algoritmo k-medias (6 grupos)

- nstar es cantidad de subconjuntos aleatorios que se escogen para realizar los calculos de algoritmo.
- el 6 es el nmero de clouster o de agrpupaciones.

```
Kmeans.6<-kmeans(X.s, 6, nstart=25)
```

centroides

```
Kmeans.6$centers
```

```
##   Log-Population     Income Log-Illiteracy   Life Exp     Murder     HS Grad
## 1    -1.65470747  2.1094604     -0.3490974 -1.2728011  1.0895183  1.58994719
## 2     0.12233125 -1.3014617      1.3019262 -1.1773136  1.0919809 -1.41578257
## 3    -1.30355300 -0.2681986     -0.9775813  0.3548885 -0.9218376  0.46019574
```

```
## 4    -0.02012796  0.2632441     -1.0527537  1.1656294 -0.9511840  0.92206977
## 5    -0.15758822  0.9109826      0.2165582  0.5182427 -0.6480455  0.18472210
## 6     1.05203572  0.2689748      0.1658871 -0.1124169  0.4831422 -0.06765652
##        Frost    Log-Area
## 1  1.2608490  1.51085951
## 2 -0.7206500  0.07602772
## 3  1.1526361  0.03872450
## 4  0.3010938  0.49075236
## 5 -0.1187800 -1.92526117
## 6 -0.4380016  0.37632593
```

cluster de pertenencia

```
Kmeans.6$cluster
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              2              1              6              2              6
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              4              5              5              6              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              5              3              6              6              4
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              4              2              2              3              5
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              5              6              4              2              6
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              4              1              3              5
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              6              2              3              6
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              6              4              6              5              2
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              2              6              4              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              6              4              2              4              3
```

4.- SCDG

```
SCDG<-sum(Kmeans.6$withinss)
SCDG
```

```
## [1] 121.0769
```

5.- Clusters

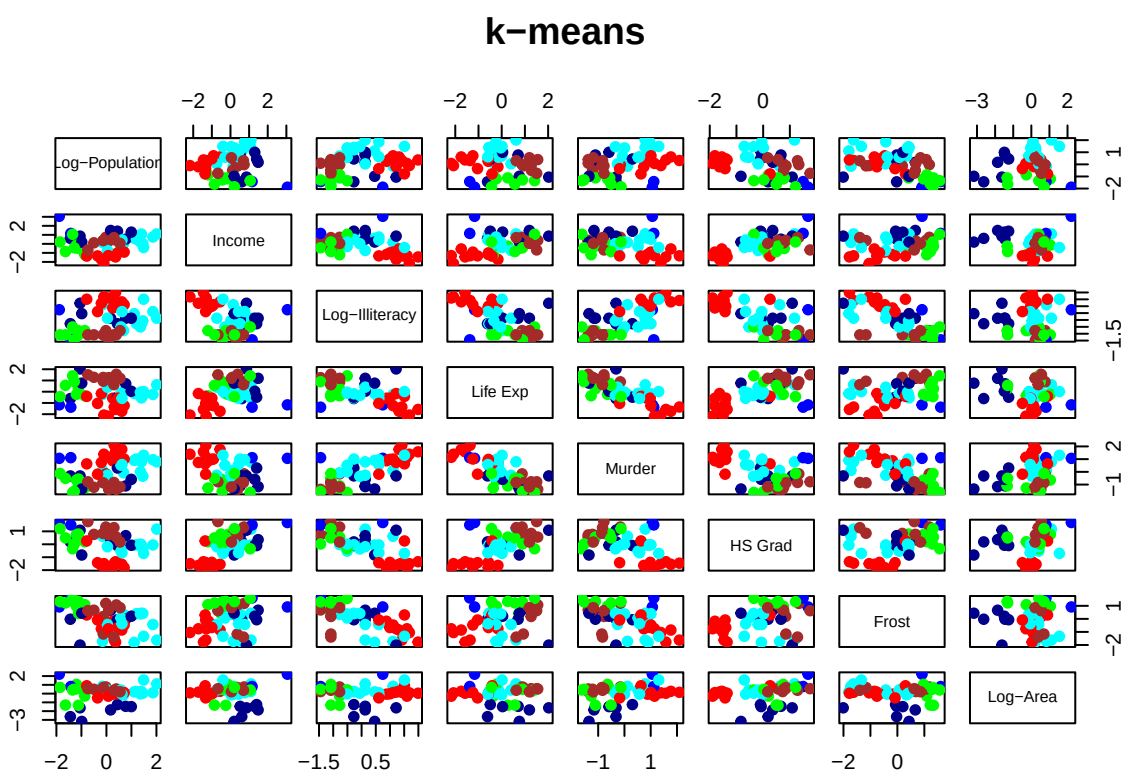
```
cl.kmeans<-Kmeans.6$cluster
cl.kmeans
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              2              1              6              2              6
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              4              5              5              6              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              5              3              6              6              4
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              4              2              2              3              5
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              5              6              4              2              6
```

```
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              4              1              3              5
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              6              2              3              6
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              6              4              6              5              2
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              2              6              4              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              6              4              2              4              3
```

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

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

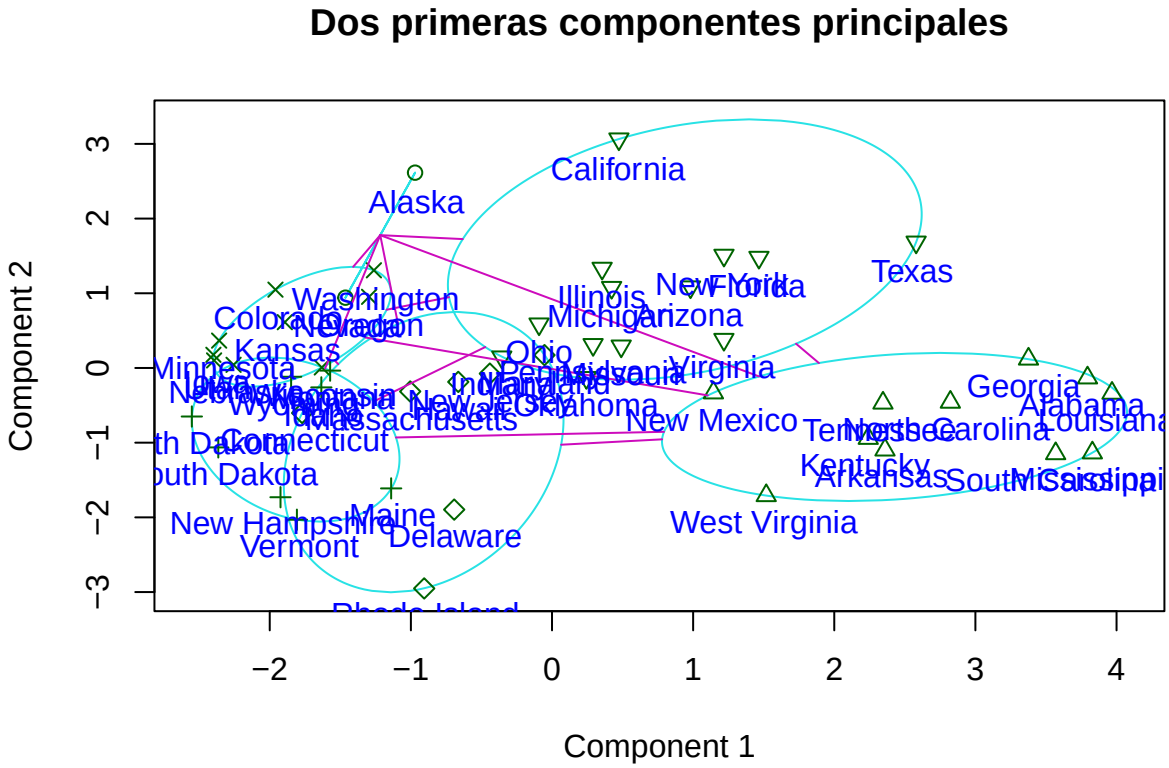



## Visualizacion con las dos componentes principales

```
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.

## Silhouette

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

1.- Generacion de los calculos

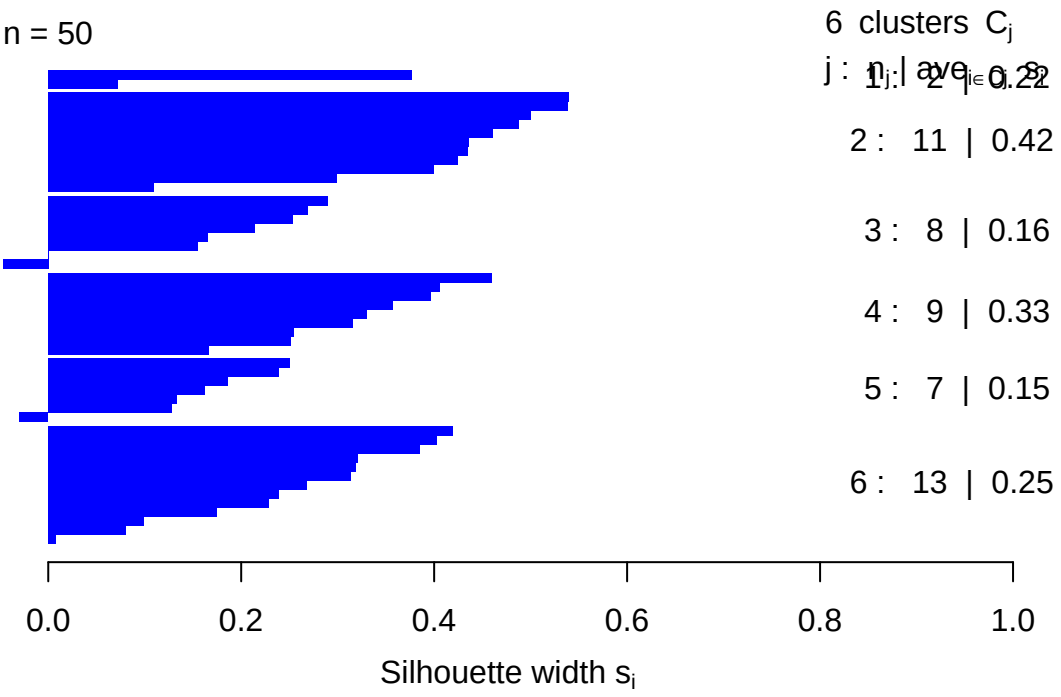
```
dist.Euc<-dist(X.s, method = "euclidean")
```

el cl.kmeans es dode se se encuentran los closters

```
Sil.kmeans<-silhouette(cl.kmeans, dist.Euc)
```

2.- Generacion del grafico

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




se utilizo un nuevo numero de clousters en este caso fueron 6, y se disminuyo significativamente la suma de cuadrados dentro del grupo pero la probailidad de agrupamiento es muy baja para la mayoria de los grupos, el unico mas significativo es 6 y 4 no es probabilidad, es el ancho de silhouet el promedio de siluedt debe ser alto, en este caso es de 0.27 por lo que se debe buscar otro numero de clousters se recomienda bajar el numero de clouster a 2 y volver a correr el codigo