k-means

Melissa Ortega Galarza

2022-06-02

1. Replica

Cargar la matriz de datos.

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

Transformación 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"
```

Método 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 (3 grupos) cantidad de subconjuntos aleatorios que se escogen para realizar los calculos 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     -0.7900149  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

```
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

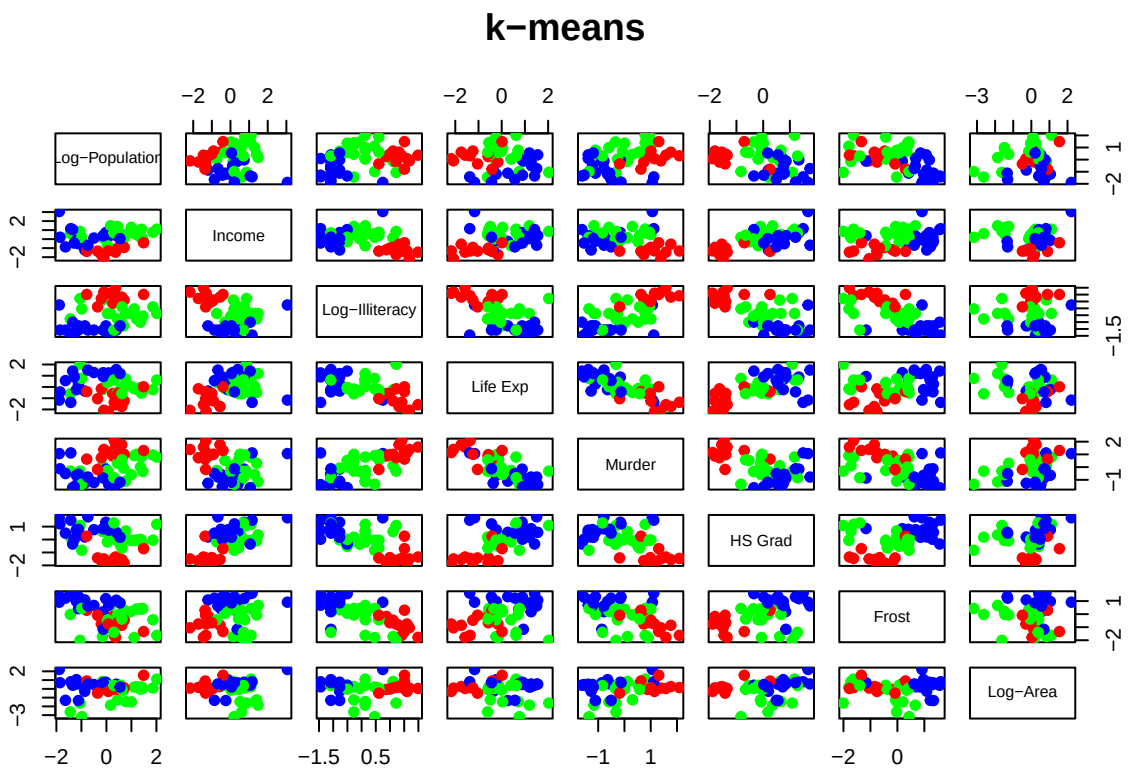
```
SCDG<-sum(Kmeans.3$withinss)
```

5.- Clusters

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

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)
```

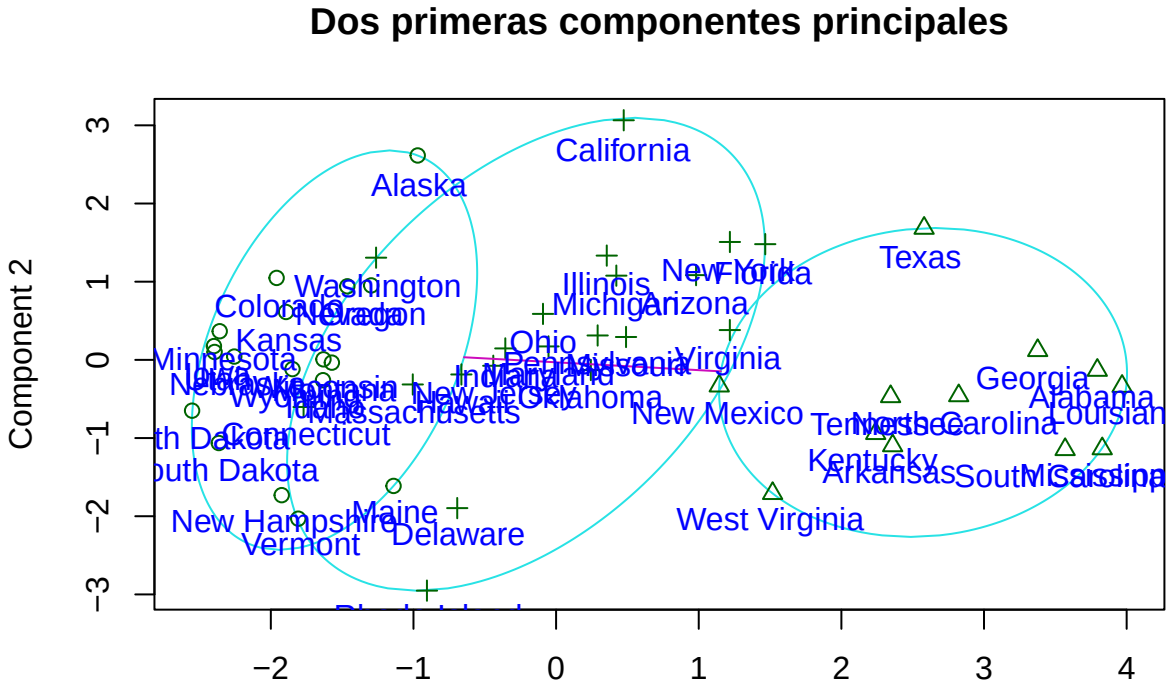


Visualización con las dos componentes principales

```
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.

Silhouette

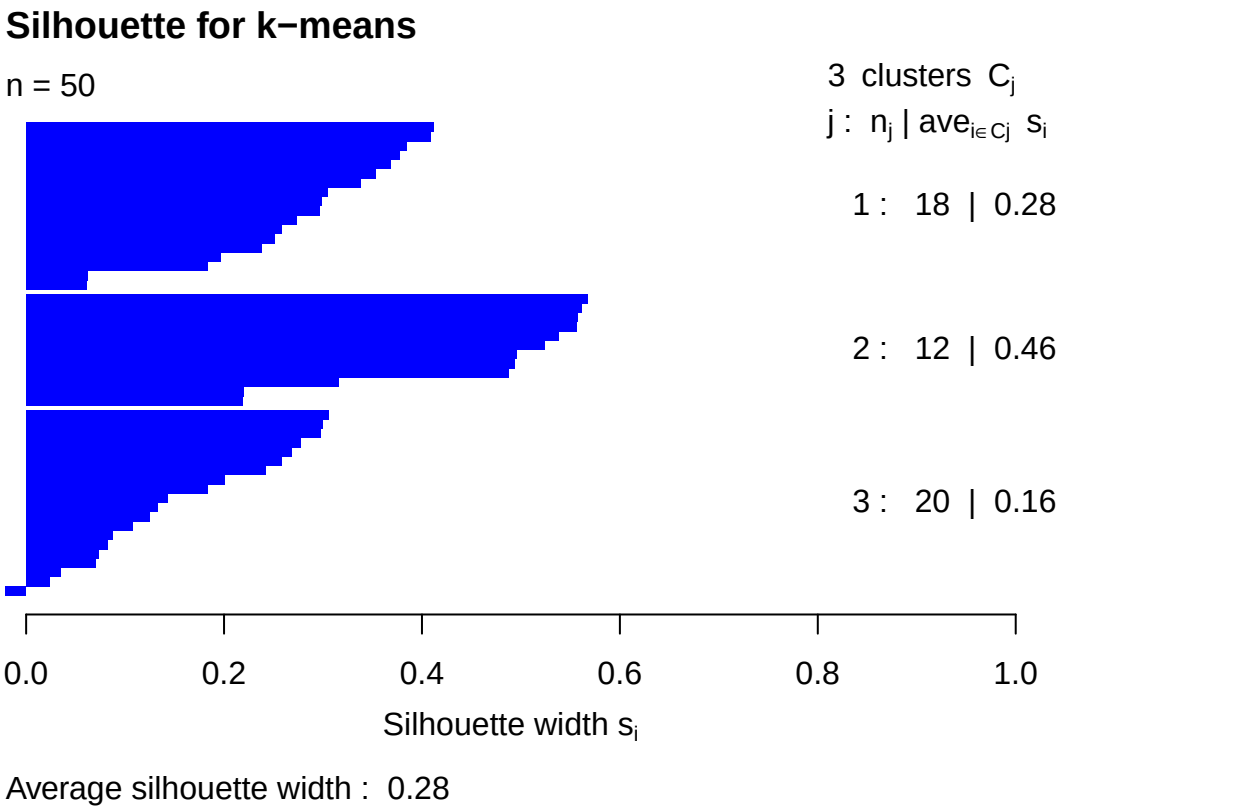
Representacion grafica de la eficacia de clasificación de una observacion dentro de un grupo.

1.- Generacion de los calculos

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

2.- Generación del gráfico

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



2.Cambio de Número de Clusters

Cargar la matriz de datos.

Aqui se consideran las medianas busca k objetos representativos

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

Transformacion 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

1.- Separación 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 (5 grupos) cantidad de subconjuntos aleatorios que se escogen para realizar los calculos de algoritmo.

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

Centroides

```
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     -0.5470524  0.0007323385     -1.0134235  0.8605152 -0.9878669  0.67299139
## 4      0.1223312 -1.3014616989      1.3019262 -1.1773136  1.0919809 -1.41578257
## 5     -1.7220507  1.4769369102     -0.5929507 -0.9946909  0.6831838  1.46407534
##        Frost    Log-Area
## 1 -0.1187800 -1.92526117
## 2 -0.4380016  0.37632593
## 3  0.6632731  0.25141793
## 4 -0.7206500  0.07602772
## 5  1.2800868  1.24186646
```

Cluster de pertenencia

```
Kmeans.5$cluster
```

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

4.- SCDG

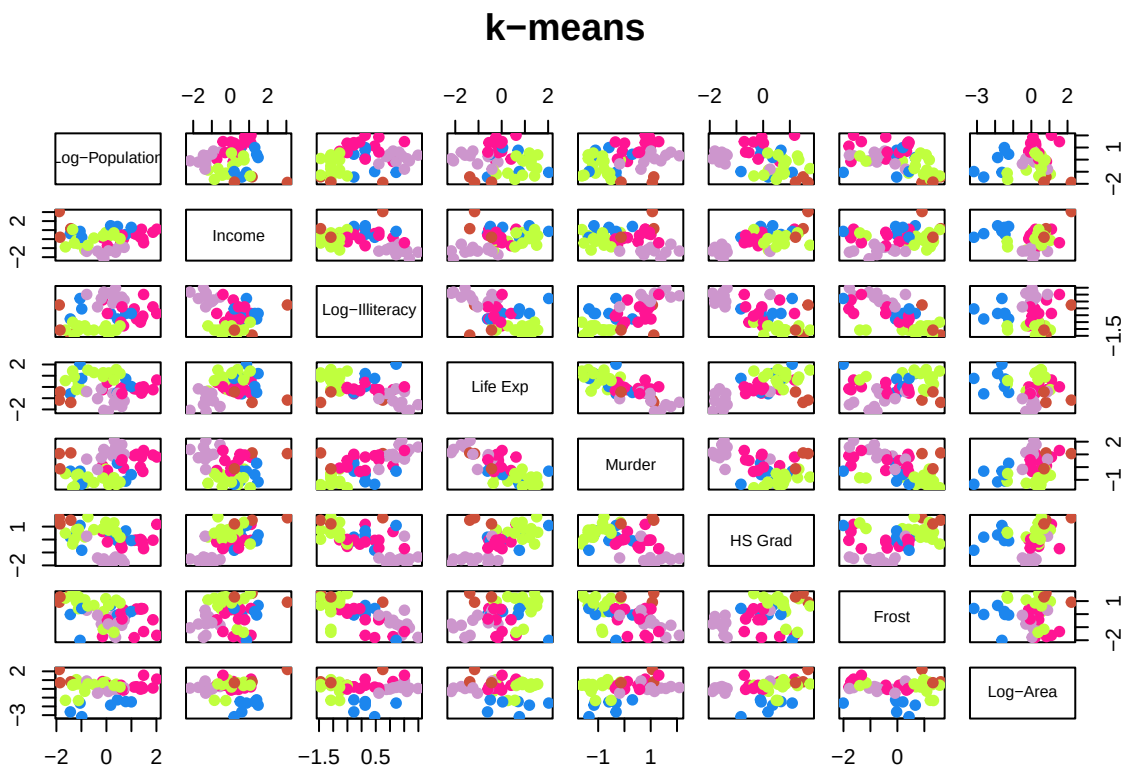
```
SCDG<-sum(Kmeans.5$withinss)
```

5.- Clusters

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

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

```
col.cluster<-c("dodgerblue2", "deeppink", "olivedrab1","plum3","tomato3")[cl.kmeans]
pairs(X.s, col=col.cluster, main="k-means", pch=19)
```

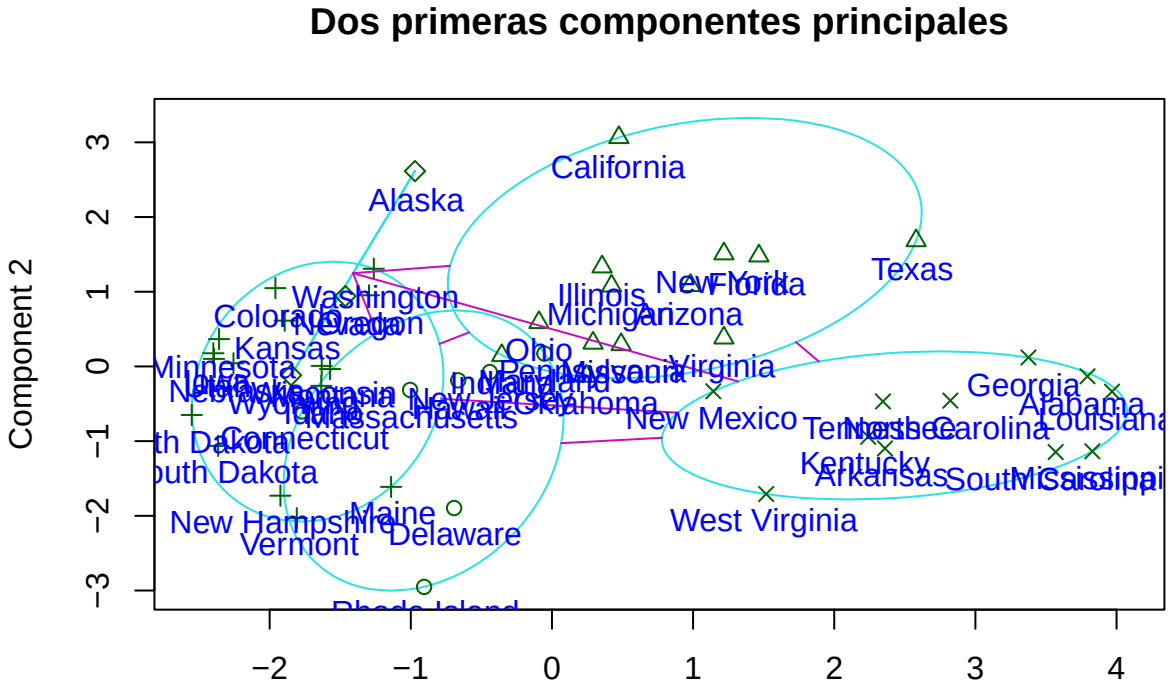


Visualización con las dos componentes principales

```
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.

Silhouette

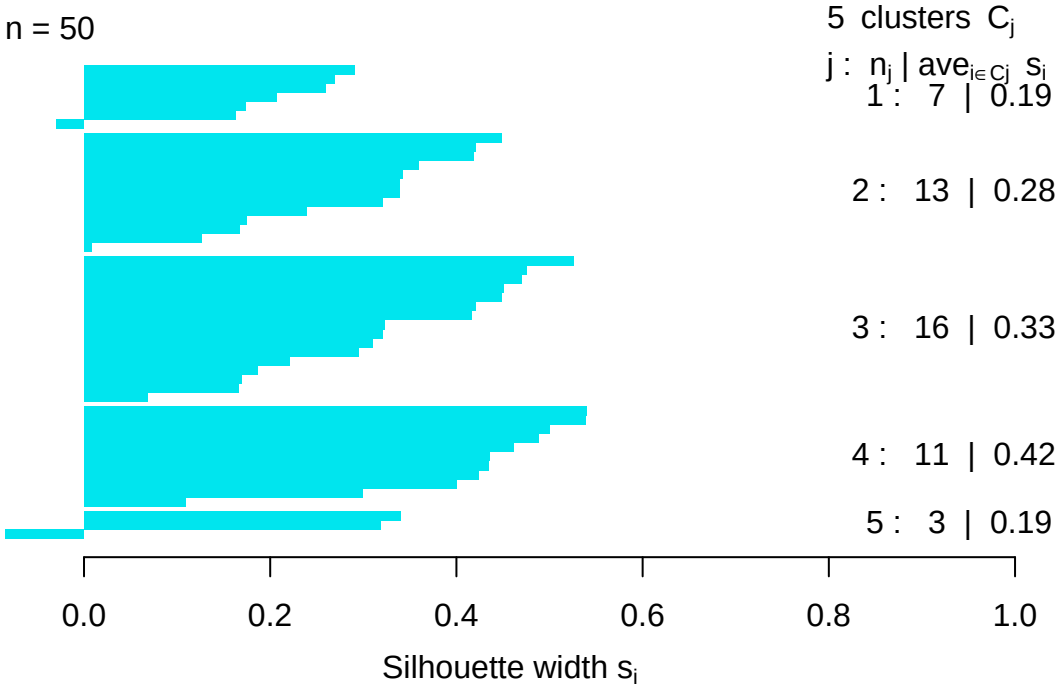
Representacion grafica de la eficacia de clasificación de una observacion dentro de un grupo.

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="turquoise2")
```



Análisis:

Se utilizo un nuevo numero de clousters en este caso fueron 5, 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 3 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 busacar otro numero de clousters Se da como consejo bajar el numero de clouster a 4 y volver a replicar el codigo