# K-means

Itzel Teodocio Olivares

2022-06-05

#_____ 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.

```
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)
SCDG
```

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

5.- Clusters

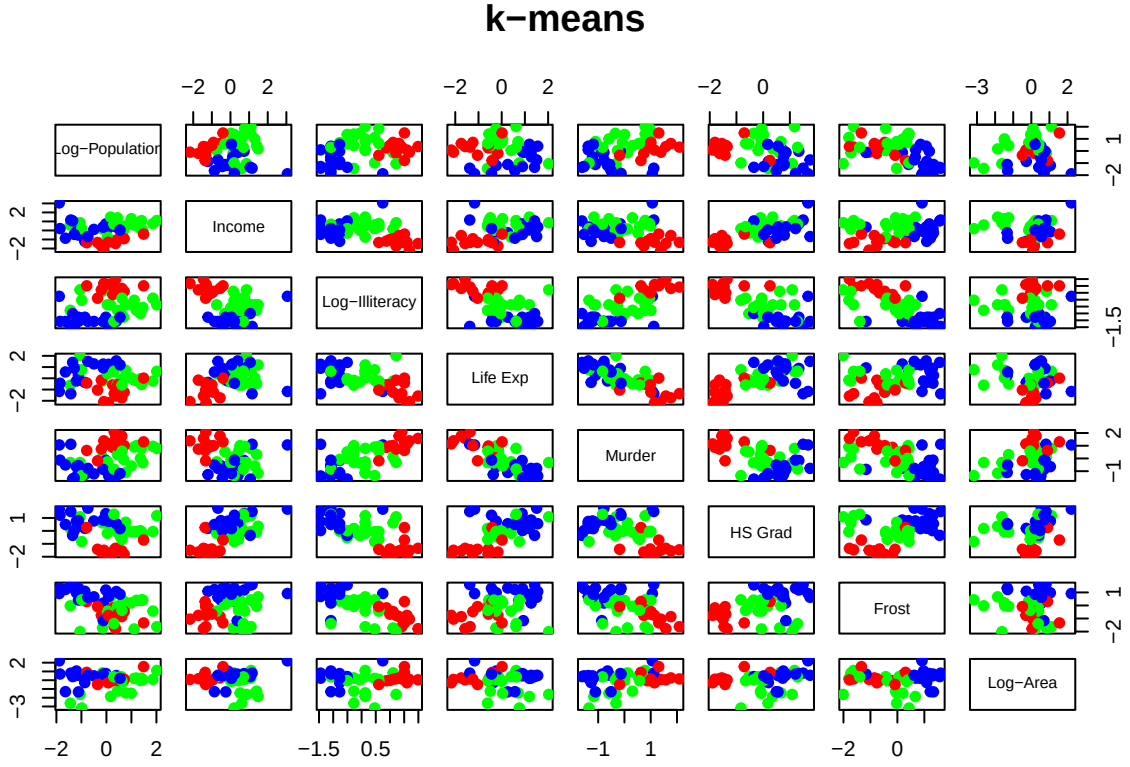
```
cl.kmeans<-Kmeans.3$cluster
cl.kmeans
```

```
##        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)
```

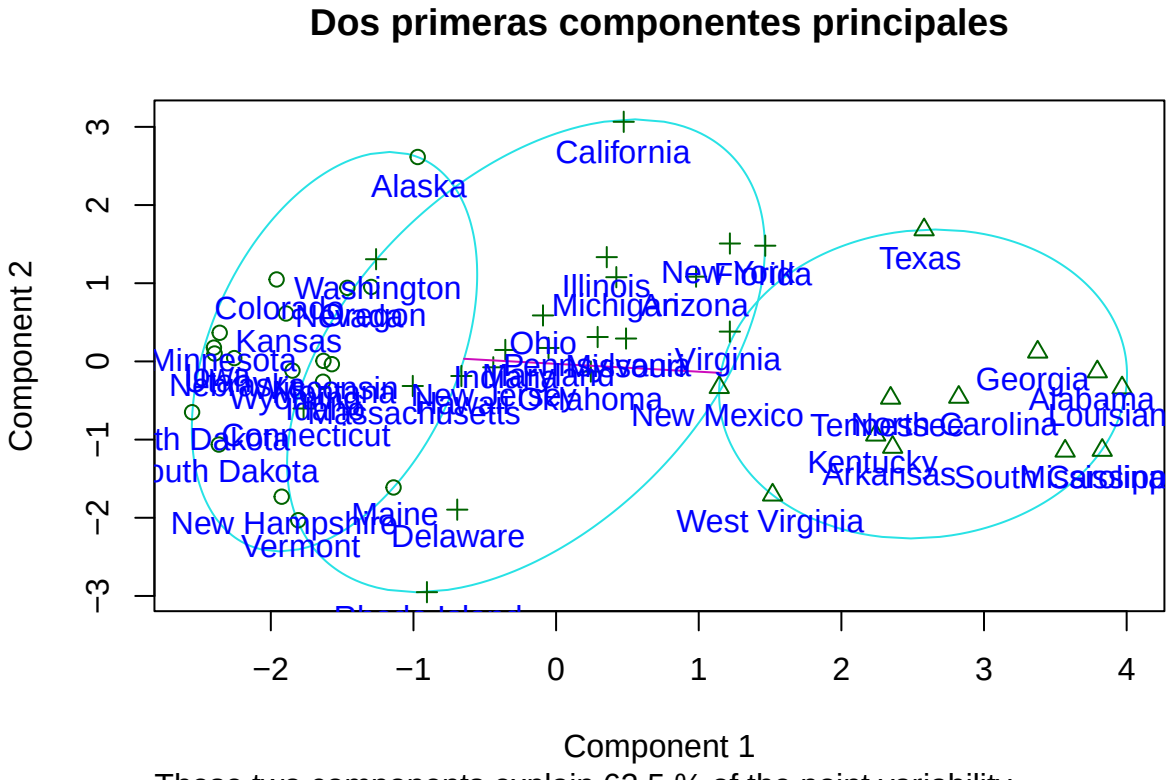



# Visualizacion con las dos componentes principales

```
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

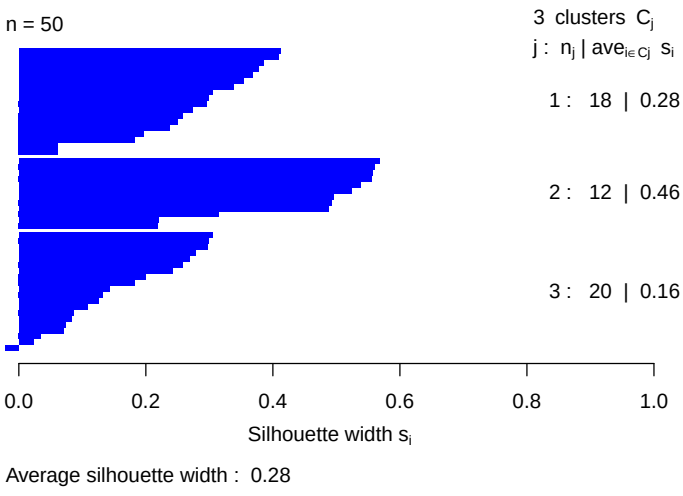
# Representacion grafica de la eficacia de # clasificacion de una observacion dentro de un grupo.

1.- Generacion de los calculos

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

2.- Generacion del grafico

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



## Sugerir nuevo número de clusters

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.

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

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

```
Kmeans.3<-kmeans(X.s, 2, nstart=25)
```

centroides

```
Kmeans.3$centers
```

```
##   Log-Population     Income Log-Illiteracy   Life Exp     Murder    HS Grad
## 1      0.3921592 -0.7973132      1.1635825 -0.8863645  0.9913208 -1.0270524
## 2     -0.1845455  0.3752062     -0.5475682  0.4171127 -0.4665039  0.4833188
##        Frost   Log-Area
## 1 -0.8493032  0.2164565
## 2  0.3996721 -0.1018619
```

cluster de pertenencia

```
Kmeans.3$cluster
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              2              1              1              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              2              2              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              2              2              2              2
```

```
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              1              1              2              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              2              2              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              2              2              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              2              2
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              2              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              1              1              2              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              2              1              2              2
```

4.- SCDG

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

```
## [1] 257.0639
```

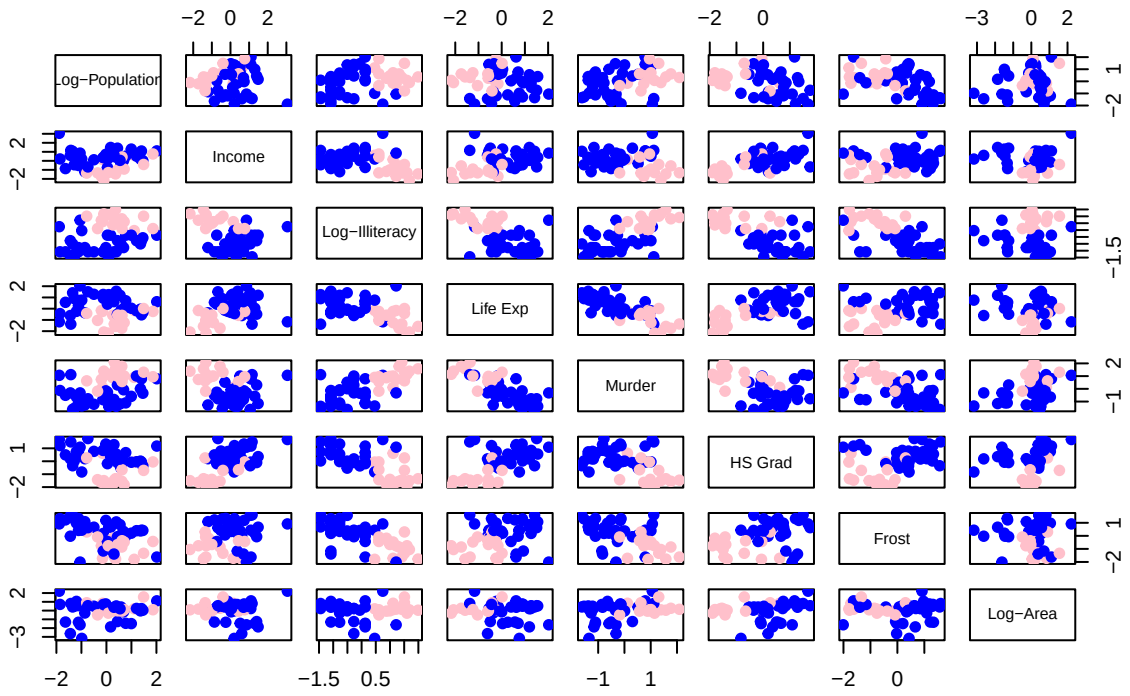
5.- Clusters

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

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

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

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

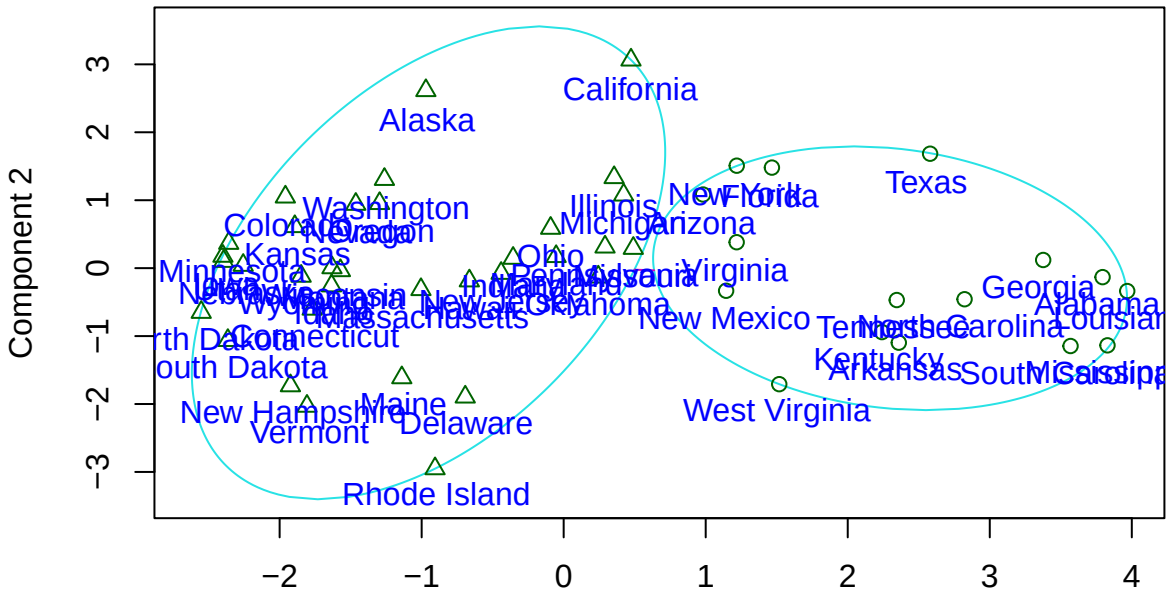



# Visualizacion con las dos componentes principales

```
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

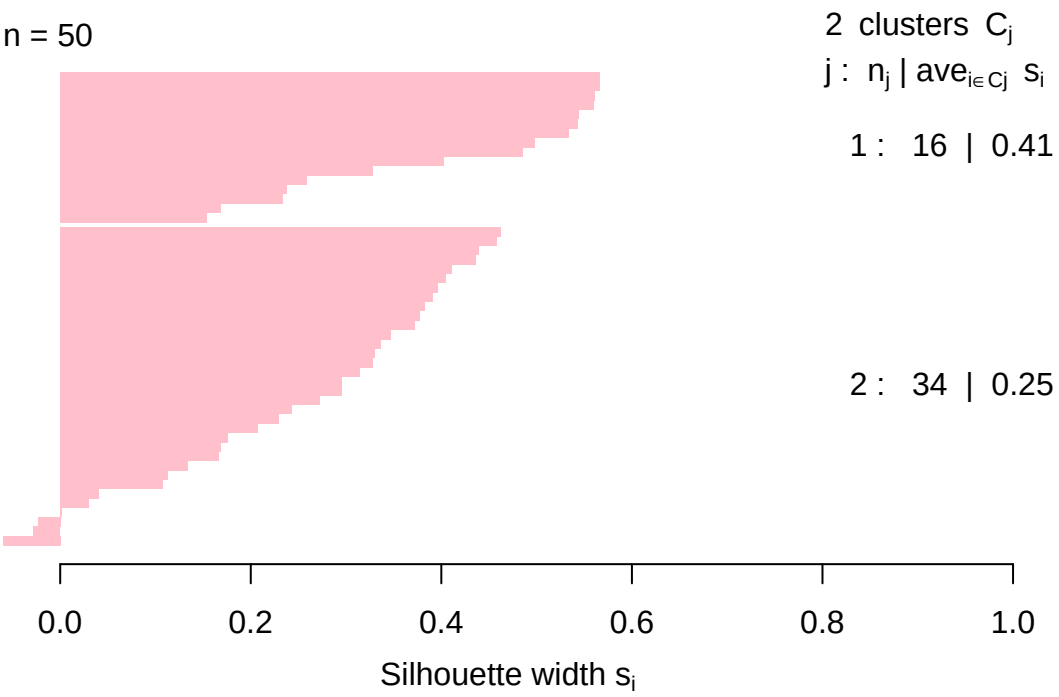
#———————————— Representacion grafica de la eficacia de clasificacion de una observacion dentro de un grupo.

1.- Generacion de los calculos

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

2.- Generacion del grafico

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




Average silhouette width : 0.3