K-means

Angel Caballero Domimguez

5/1/2022

Estudio 2: Clustering K-means

Paquetes

Vamos a preparar los paquetes que vamos a utilizar:

```
# Paquete para visualización de los clusters
#install.packages("factoextra")
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
# Paquete para la silueta de los clusters
#install.packages("cluster")
library(cluster)
```

Lectura de los datos

```
df<-read.csv("./Prep_StudentsPerformance.csv",stringsAsFactors = TRUE, header = TRUE)
```

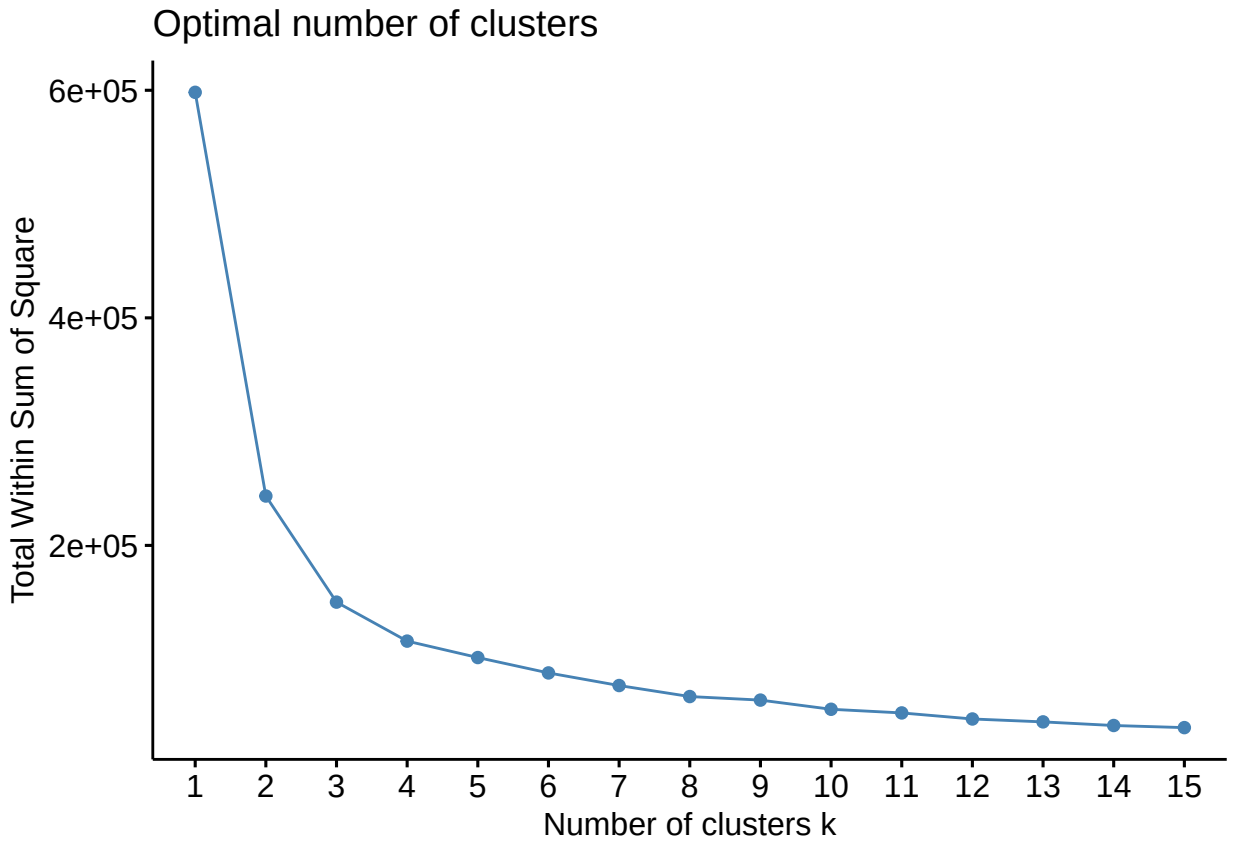
Extraemos las puntuaciones de los tests y las guardamos por separado.

```
scores <- df[,6:8]
```

Número de clusters óptimo

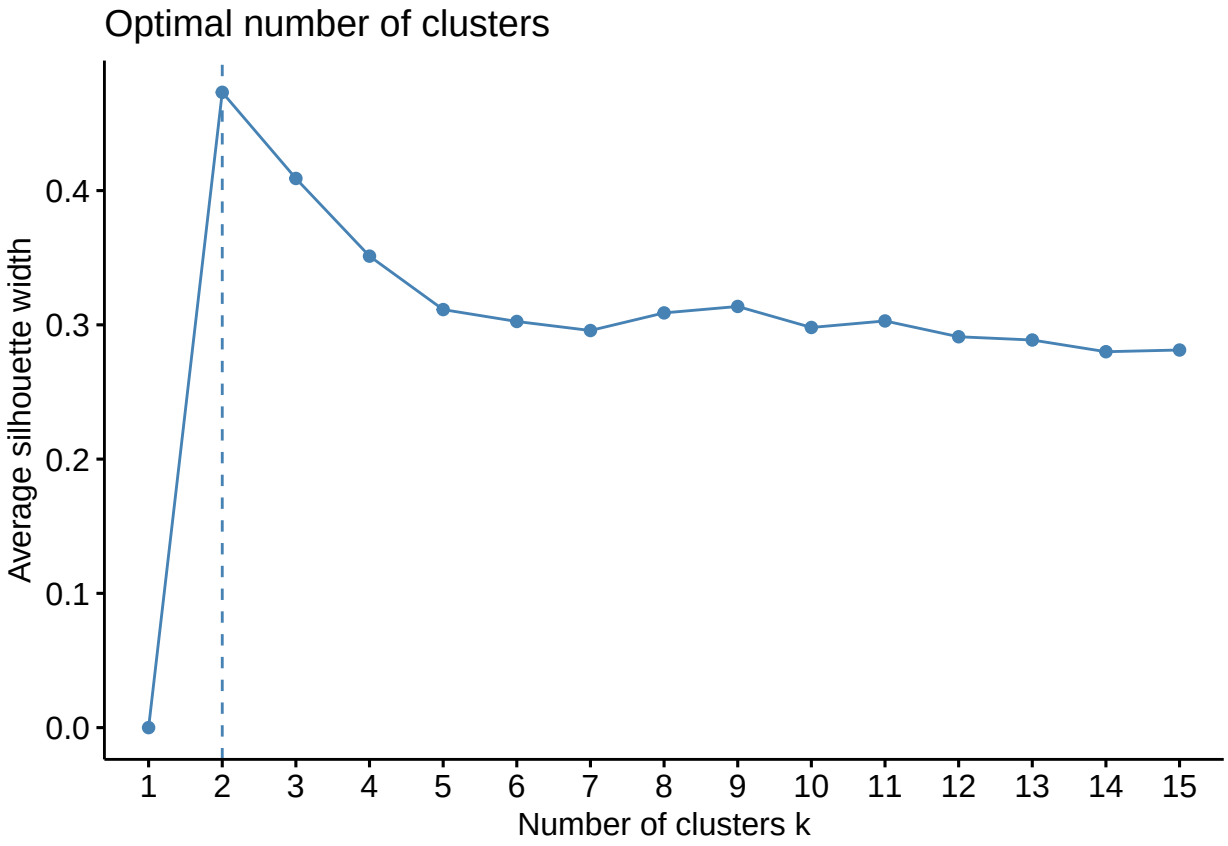
Vamos a buscar cuál es el número de clústers óptimo para nuestro caso. Y para ello, primero vamos a ver la compactación del data set según la suma de cuadrados del clúster (SSE).

```
fviz_nbclust(scores, kmeans, method = "wss", k.max = 15)
```



Como se puede observar, un buen número de clusters podrían ser 2. Para confirmar esto, vamos a ver los resultados de la silueta.

```
fviz_nbclust(scores, kmeans, method = "silhouette", k.max = 15)
```



Debido a los resultados obtenidos en ambos gráficos, vamos a clasificar en dos clusters.

Realización de k-means

Primero debemos establecer una semilla, para que el experimento pueda ser recreado.

```
set.seed(6)
```

Ahora vamos a utilizar la función kmeans() del paquete stasts.

```
km_res <- kmeans(scores,centers =2,nstart =20)

# Mostramos un resumen del resultado
summary(km_res)
```

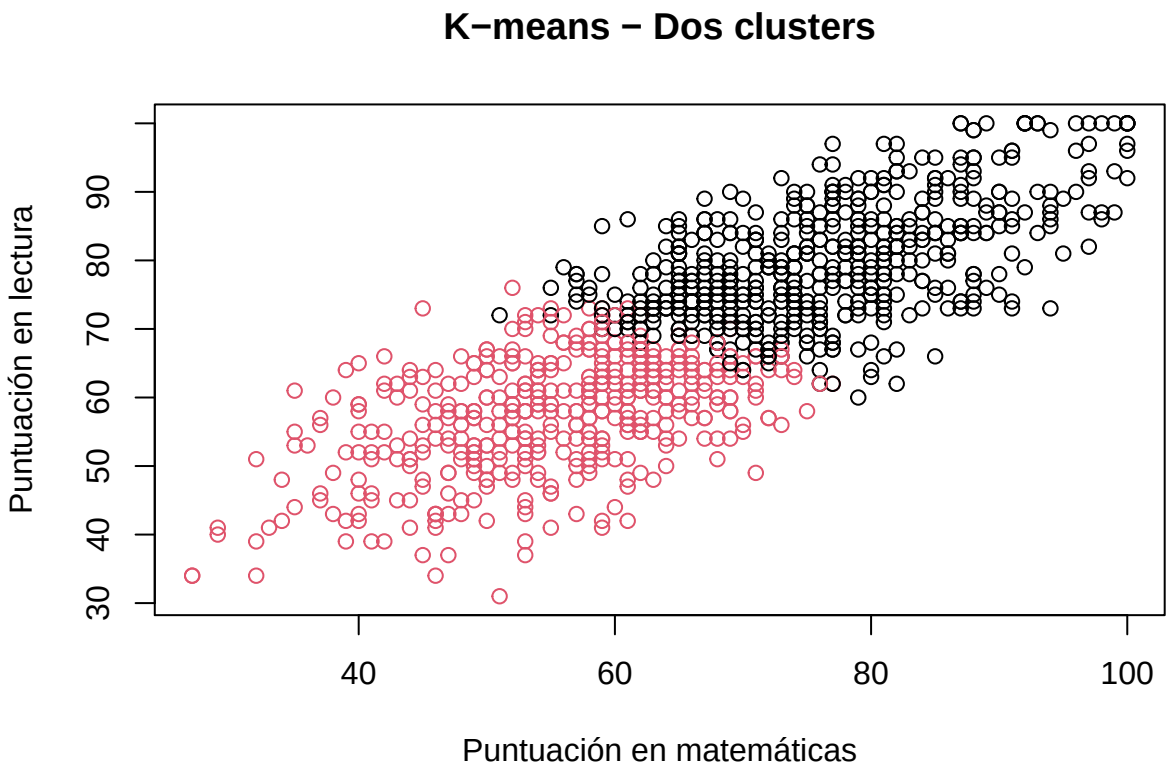
```
##              Length Class  Mode
## cluster      986    -none- numeric
## centers        6    -none- numeric
## totss          1    -none- numeric
## withinss       2    -none- numeric
## tot.withinss   1    -none- numeric
## betweenss      1    -none- numeric
## size           2    -none- numeric
## iter           1    -none- numeric
## ifault         1    -none- numeric
```

```r
# Mostramos todo el contenido del objeto resultante
print(km_res)
```

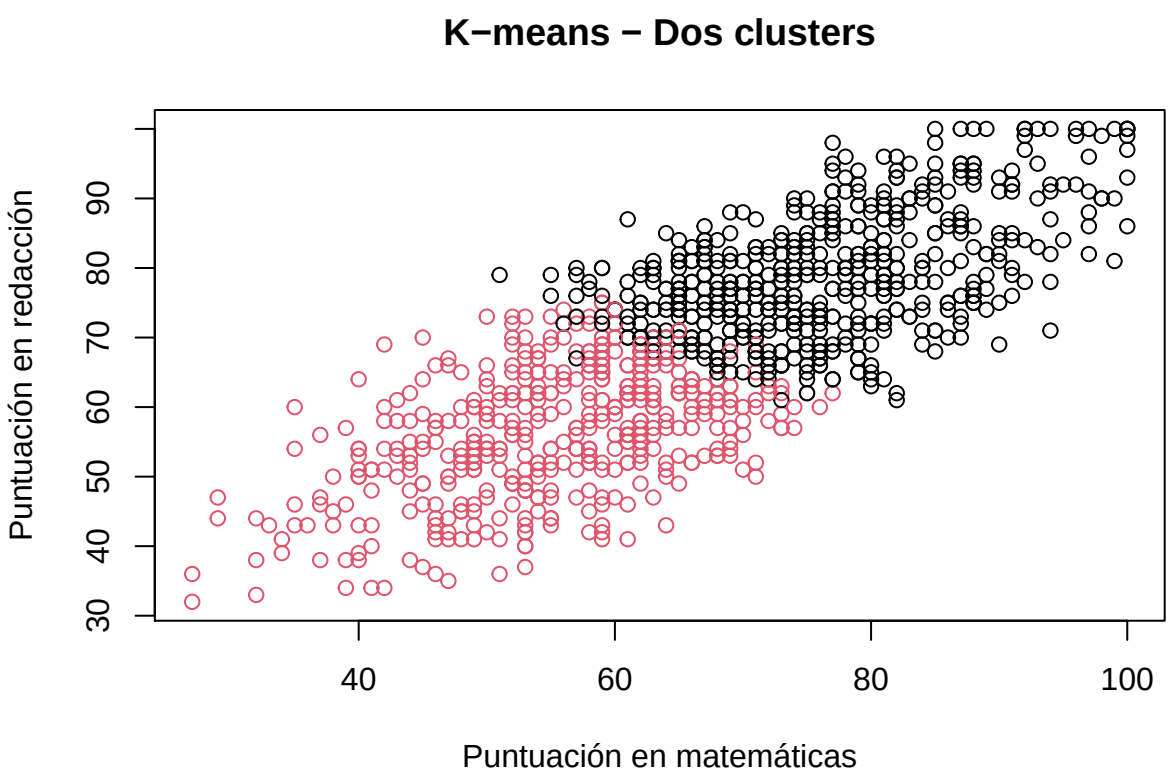
```
## K-means clustering with 2 clusters of sizes 536, 450
##
## Cluster means:
##       Math  Reading  Writing
## 1 76.34515 79.81530 78.99813
## 2 55.20000 57.70222 56.32000
##
## Clustering vector:
##   [1] 1 1 1 2 1 1 1 2 2 2 2 2 1 1 2 1 1 2 2 2 1 2 1 1 1 2 1 1 1 1 2 2 2 1 1 1 2
##  [38] 1 2 2 2 2 2 2 2 2 2 1 1 1 2 1 2 1 1 2 1 2 2 1 2 2 1 2 2 2 1 2 2 2 2 2 2 2 2
##  [75] 1 1 2 2 2 2 2 2 1 1 1 2 1 1 2 1 2 1 1 2 1 2 2 1 1 1 2 1 2 1 2 2 1 1 2 2 2
## [112] 1 1 1 1 2 2 1 1 1 2 1 1 1 1 1 2 1 2 1 1 1 2 2 2 2 2 1 2 2 2 1 1 1 1 1 2 1
## [149] 2 2 2 1 1 2 1 2 1 1 2 2 1 1 2 1 1 1 1 1 1 2 2 1 2 1 2 1 1 2 2 1 2 2 1 2 2
## [186] 1 2 1 2 2 1 2 2 2 2 1 1 1 1 2 2 1 1 1 1 2 1 2 2 1 1 1 2 1 2 2 1 2 1 2 2 1
## [223] 2 1 1 1 2 1 1 1 1 2 2 2 1 1 1 2 2 1 1 1 2 2 2 2 1 2 1 2 1 1 1 1 1 1 1 2 1
## [260] 1 2 1 1 1 1 2 2 2 1 1 1 1 2 1 2 2 2 1 1 2 1 1 1 1 1 1 1 2 1 1 2 2 1 2 1 1
## [297] 2 1 1 1 1 1 2 2 2 1 2 2 1 2 2 1 1 1 2 1 1 1 2 2 1 2 1 2 2 2 2 1 1 2 1 2 2
## [334] 2 2 1 1 1 1 2 1 1 1 2 2 1 2 2 1 2 2 2 1 2 1 2 2 2 2 2 2 2 1 1 2 1 1 1 2 1
## [371] 1 1 2 1 1 1 2 2 1 1 1 2 1 2 1 1 2 1 2 2 1 2 2 2 2 2 1 2 1 2 1 2 1 1 1 2 2
## [408] 1 1 1 1 2 2 1 2 2 1 2 2 1 1 2 2 2 1 2 2 1 2 1 1 2 1 1 1 2 1 1 1 2 1 2 1 1
## [445] 1 1 2 2 2 1 2 1 1 2 2 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 2 2 2 2 1 2 1 2
## [482] 1 1 2 1 1 2 2 2 1 1 1 1 1 2 1 2 1 1 2 1 1 1 2 2 2 1 1 1 1 1 1 2 1 2 2 2 2
## [519] 2 2 2 1 1 2 2 1 1 1 2 2 1 1 1 1 1 1 2 1 1 1 2 1 1 1 2 1 2 2 1 2 2 2 1 1 1
## [556] 1 2 2 1 1 2 1 1 1 2 2 1 2 2 1 2 2 1 1 1 1 1 1 2 2 2 2 2 2 2 2 1 1 1 2 1 1 2
## [593] 1 2 1 1 1 2 2 2 2 2 1 1 1 2 2 1 1 1 2 2 2 1 2 1 2 2 2 2 2 2 1 1 1 1 1 1 1 1
## [630] 1 2 1 1 1 2 1 2 2 2 1 2 1 1 1 1 2 1 2 2 1 1 2 2 2 1 2 1 1 1 1 1 2 1 1 1 2
## [667] 1 1 1 2 1 2 2 2 2 1 1 1 2 1 2 1 1 1 2 1 1 2 1 2 1 2 1 2 2 2 2 2 1 2 1 1 1
## [704] 1 1 1 1 1 1 1 1 2 1 2 2 1 1 2 1 2 1 2 1 2 2 2 1 2 1 2 1 2 1 1 2 1 1 2 2 1
## [741] 1 1 1 1 2 1 2 1 2 1 2 2 1 2 2 1 1 2 1 2 2 1 2 1 2 2 1 2 1 1 2 1 1 2 1 2 1
## [778] 2 2 2 2 1 1 2 2 1 1 2 2 1 1 1 1 1 1 1 2 1 2 2 2 2 1 1 1 2 1 2 1 1 1 2 1 2
## [815] 2 2 1 1 2 2 1 2 1 2 2 2 1 2 1 2 2 1 2 1 1 2 2 1 1 2 1 1 2 1 1 1 2 1 2 1 2
## [852] 1 1 1 2 2 1 2 2 1 1 1 2 1 1 2 1 1 2 1 1 2 2 1 1 2 1 2 1 1 2 1 2 2 1 2 1 1
## [889] 1 2 1 1 1 2 1 1 1 2 1 2 2 2 1 1 2 1 1 1 2 1 2 1 2 2 2 2 2 1 2 1 1 1 2 2 2
## [926] 1 1 1 1 1 2 2 2 1 2 2 1 1 1 1 2 2 2 1 1 2 1 2 2 1 1 2 1 2 2 1 1 1 1 2 2 2
## [963] 1 2 2 2 1 1 1 1 1 2 2 1 2 1 1 1 1 1 2 1 2 2 1 1
##
## Within cluster sum of squares by cluster:
## [1] 133624.9 109768.0
##  (between_SS / total_SS =  59.3 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

A continuación, vamos a ver gráficamente los clusters resultantes por cada pareja de test.

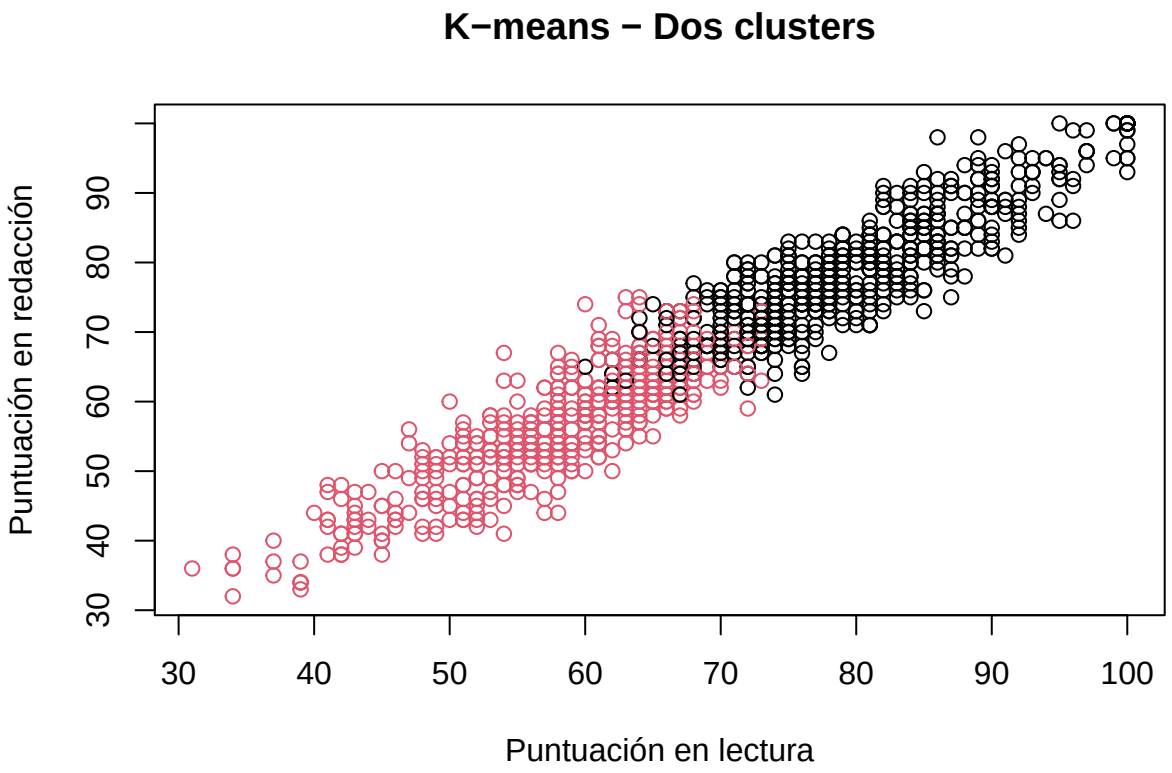
```r
# Clusters en matemáticas y lectura
plot(scores$Math,scores$Reading,col=km_res$cluster, main="K-means - Dos clusters",
     xlab = "Puntuación en matemáticas", ylab = "Puntuación en lectura")
```




```
# Clusters en matemáticas y redacción
plot(scores$Math,scores$Writing,col=km_res$cluster, main="K-means - Dos clusters",
     xlab = "Puntuación en matemáticas", ylab = "Puntuación en redacción")
```



```
# Clusters en lectura y redacción
plot(scores$Reading,scores$Writing,col=km_res$cluster, main="K-means - Dos clusters",
     xlab = "Puntuación en lectura", ylab = "Puntuación en redacción")
```



Medición de resultados

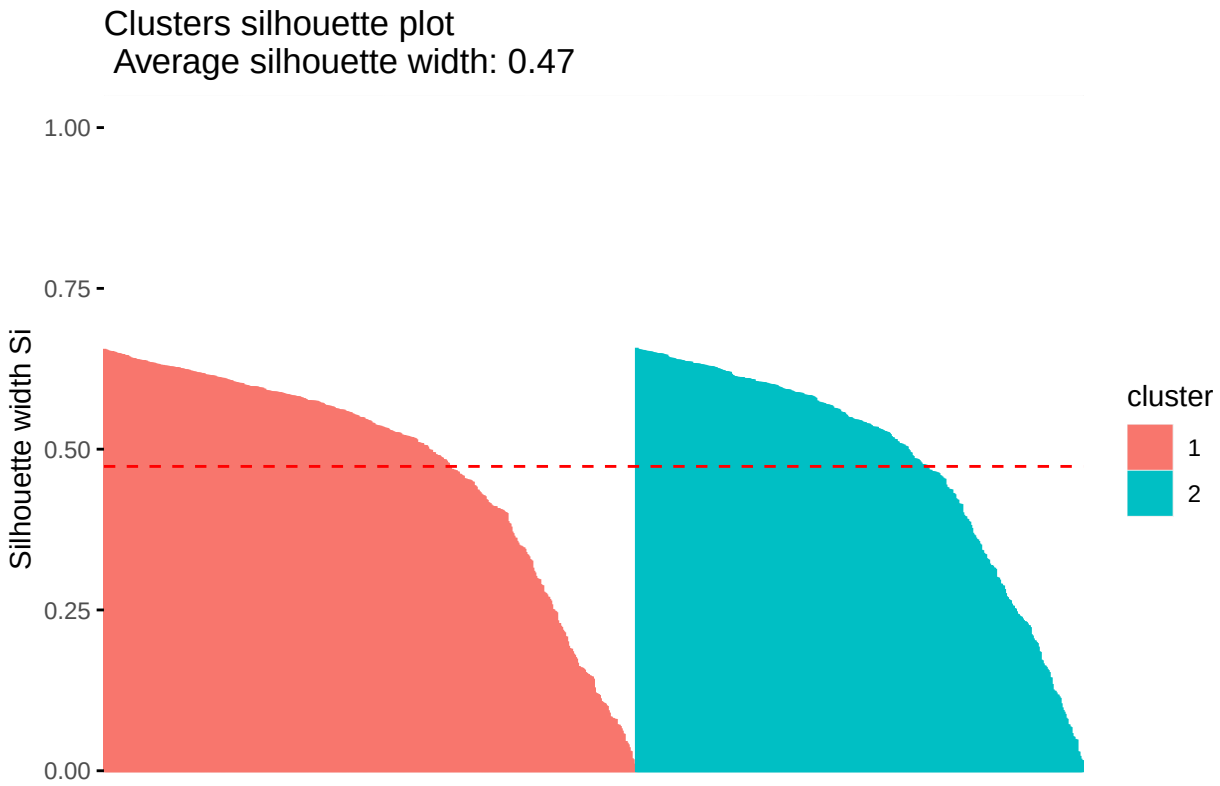
Vamos a medir la efectividad de nuestros resultados mediante la silueta de nuestros clusters.

```
sil <- silhouette(km_res$cluster, dist(scores))
head(sil[, 1:3], 10)
```

```
##       cluster neighbor sil_width
##  [1,]       1        2 0.4594115
##  [2,]       1        2 0.5656530
##  [3,]       1        2 0.5729529
##  [4,]       2        1 0.6103012
##  [5,]       1        2 0.5934719
##  [6,]       1        2 0.5952818
##  [7,]       1        2 0.5823236
##  [8,]       2        1 0.5555076
##  [9,]       2        1 0.2576756
## [10,]       2        1 0.5644183
```

```
fviz_silhouette(sil)
```

```
##   cluster size ave.sil.width
## 1       1  536          0.47
## 2       2  450          0.48
```



En el gráfico se puede observar que la silueta media es 0,47 y que la silueta tiende a 1, por lo que se han clasificado los datos correctamente. De la misma manera, no se observa ningún valor con silueta negativa, por lo que todos están clasificados en el clúster correcto.