ICO 187 ANÁLISIS DE DATOS

Clase 20: Análisis de clústers en Power Bi y funcionalidad vía web

Año: 2021
Profesor: Sebastián Egaña

1. Análisis de clústeres

1.1. ¿Qué es un clúster?

Por lo general, el concepto de asocia a la agrupación de empresas que comparten ciertas carácteristicas. Dicho concepto es aplicado también para el contexto del marketing, relacionado con la segmentación de clientes.

Veamos el siguiente set de datos:

```
utils::View(data)
```

Trabajemos con este set de datos, realizando un análisis de clústeres.

1.2. Tidy

Se deben realizar ciertas modificaciones para trabajar los datos. Por ejemplo, el algoritmo no trabaja con variables que no sean númericas. Por otra parte, también se limpian variables relacionadas con fechas.

```
data_1 <- data %>%
  select(-Date, -Month, -Weekday, - Quarter, -Day_month, -Weeknum, -Year) %>%
  mutate(Gender = ifelse(Gender == "F", 1, 0),
         Education_level = ifelse(Education_level == "Till 10th", 1,
                                  ifelse(Education_level == "Till 12th", 2,
                                         ifelse(Education_level == "Graduate", 3, 4))),
         Purpose = ifelse(Purpose == "Home", 1,
                          ifelse(Purpose == "Travel", 2,
                                 ifelse(Purpose == "Personal", 3,
                                        ifelse(Purpose == "Education", 4, 5)))))
```

1.3. Análisis de componentes principales

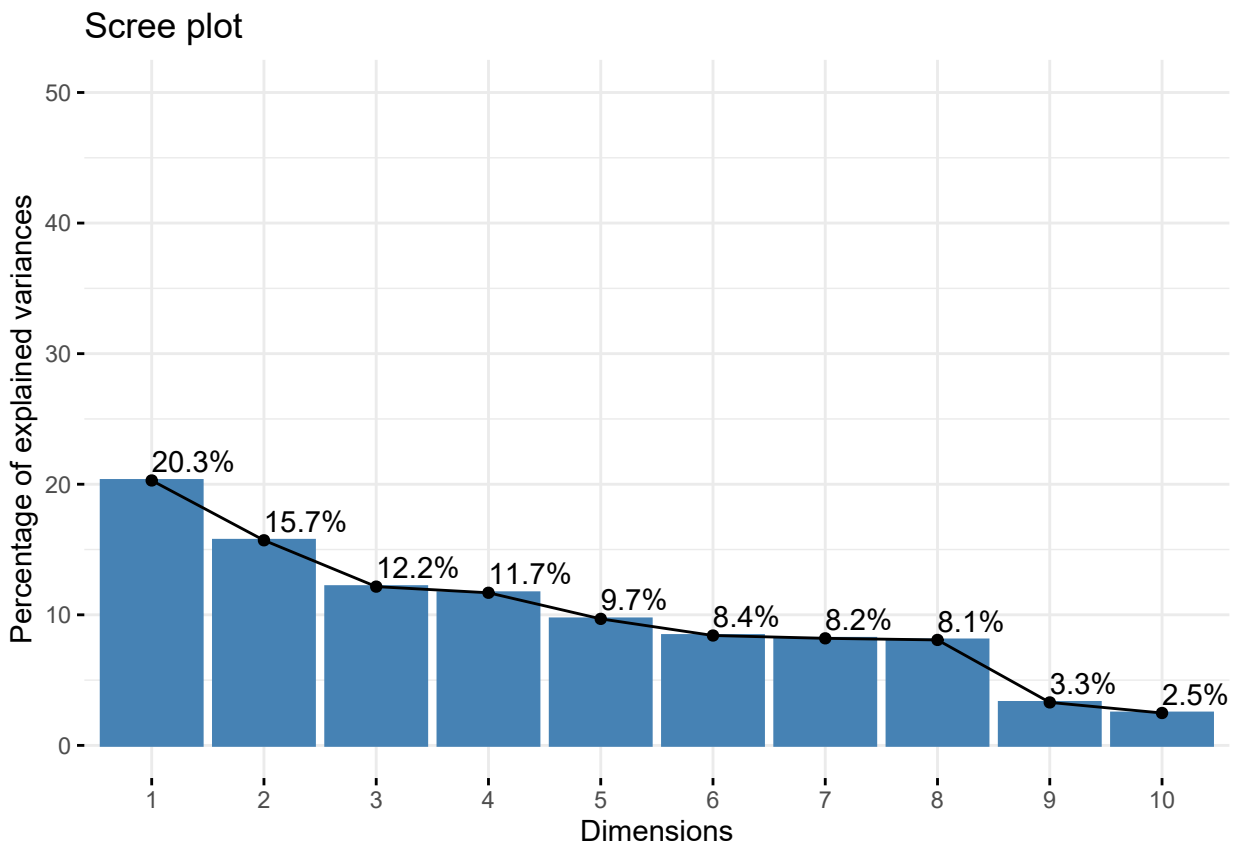
Por lo general, primero se debe determinar cuales son las variables más relevantes que explican el set de datos. Para esto, se realiza un análisis de componentes principales o principal components analysis. Veamos su implementación:

```
res.pca <- PCA(data_1,  graph = FALSE)

get_eig(res.pca)
```

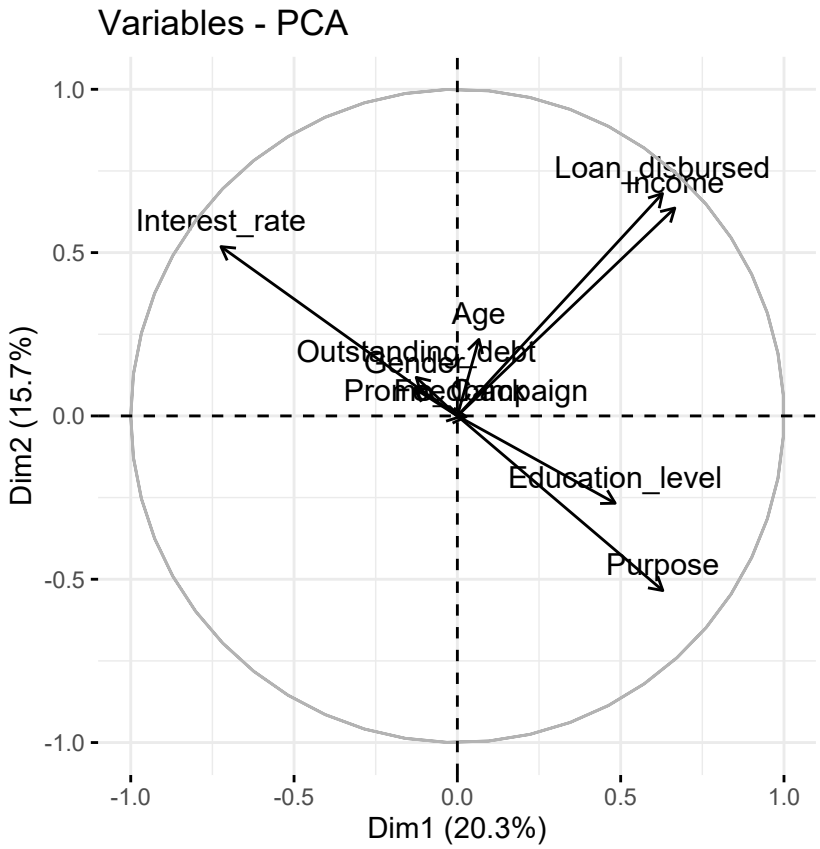
```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1   2.0288977        20.288977                    20.28898
## Dim.2   1.5707728        15.707728                    35.99671
## Dim.3   1.2159417        12.159417                    48.15612
## Dim.4   1.1688590        11.688590                    59.84471
## Dim.5   0.9689788         9.689788                    69.53450
## Dim.6   0.8415838         8.415838                    77.95034
## Dim.7   0.8199394         8.199394                    86.14973
## Dim.8   0.8077251         8.077251                    94.22698
## Dim.9   0.3291733         3.291733                    97.51872
## Dim.10  0.2481284         2.481284                   100.00000
```

```
fviz_screeplot(res.pca, addlabels = TRUE, ylim = c(0, 50))
```



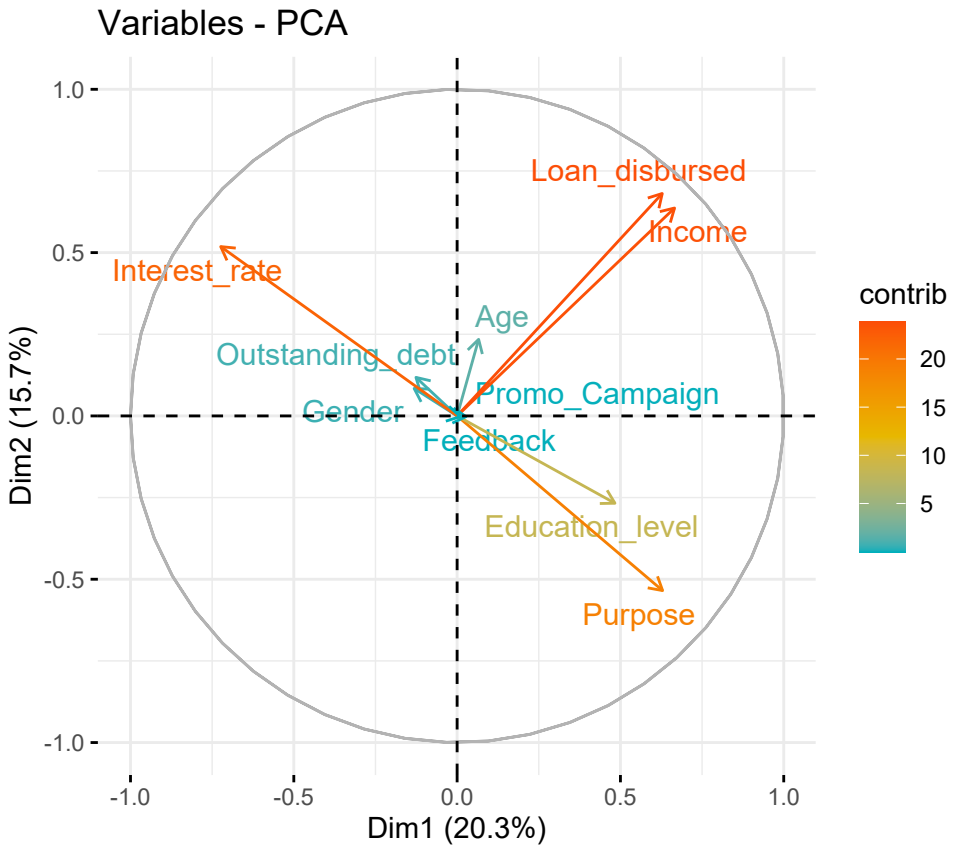
A pesar de su complejidad, solo debe entender que dicho análisis intenta explicar la variabilidad del sistema (matriz), y a la vez, establecer cuales variables son las que explican de mejor manera dicha variabilidad. Veamos el siguiente gráfico:

```
fviz_pca_var(res.pca, col.var = "black")
```



Una versión mejorada:

```
fviz_pca_var(res.pca, col.var="contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Avoid text overlapping
)
```



1.4. Análisis de clústeres

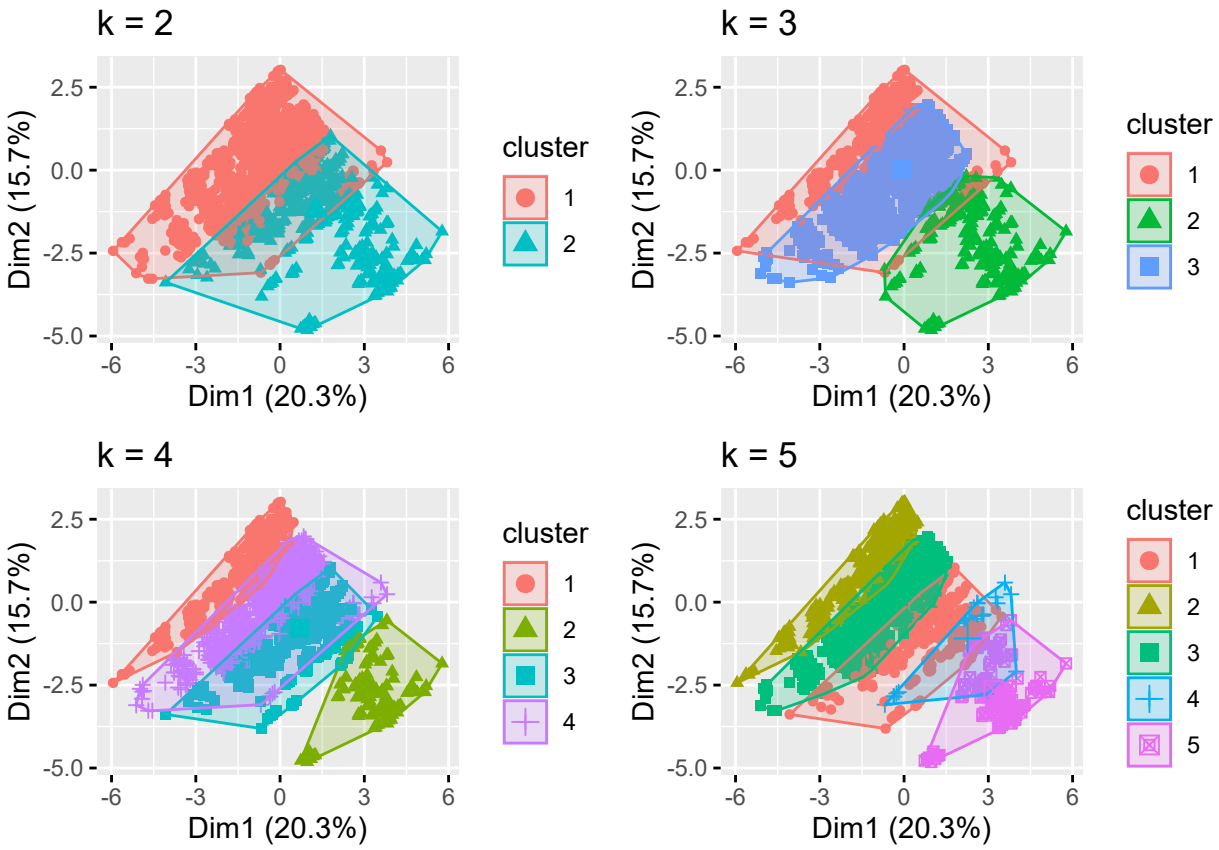
Veamos ahora la manera de construir dichos clústeres en base a los datos. Aplicaremos cuatro distintos algoritmos para evaluar la existencia de 1, 2, 3 o 4 clústeres.

```
set.seed(123)
km.res_1 <- kmeans(data_1, 2, nstart = 25)
km.res_2 <- kmeans(data_1, 3, nstart = 25)
km.res_3 <- kmeans(data_1, 4, nstart = 25)
km.res_4 <- kmeans(data_1, 5, nstart = 25)
```

Graficamos cada caso:

```
p1 <- fviz_cluster(km.res_1, geom = "point", data = data_1) +
  ggtitle("k = 2")
p2 <- fviz_cluster(km.res_2, geom = "point",  data = data_1) +
  ggtitle("k = 3")
p3 <- fviz_cluster(km.res_3, geom = "point",  data = data_1) +
  ggtitle("k = 4")
p4 <- fviz_cluster(km.res_4, geom = "point",  data = data_1) +
  ggtitle("k = 5")

grid.arrange(p1, p2, p3, p4, nrow = 2)
```

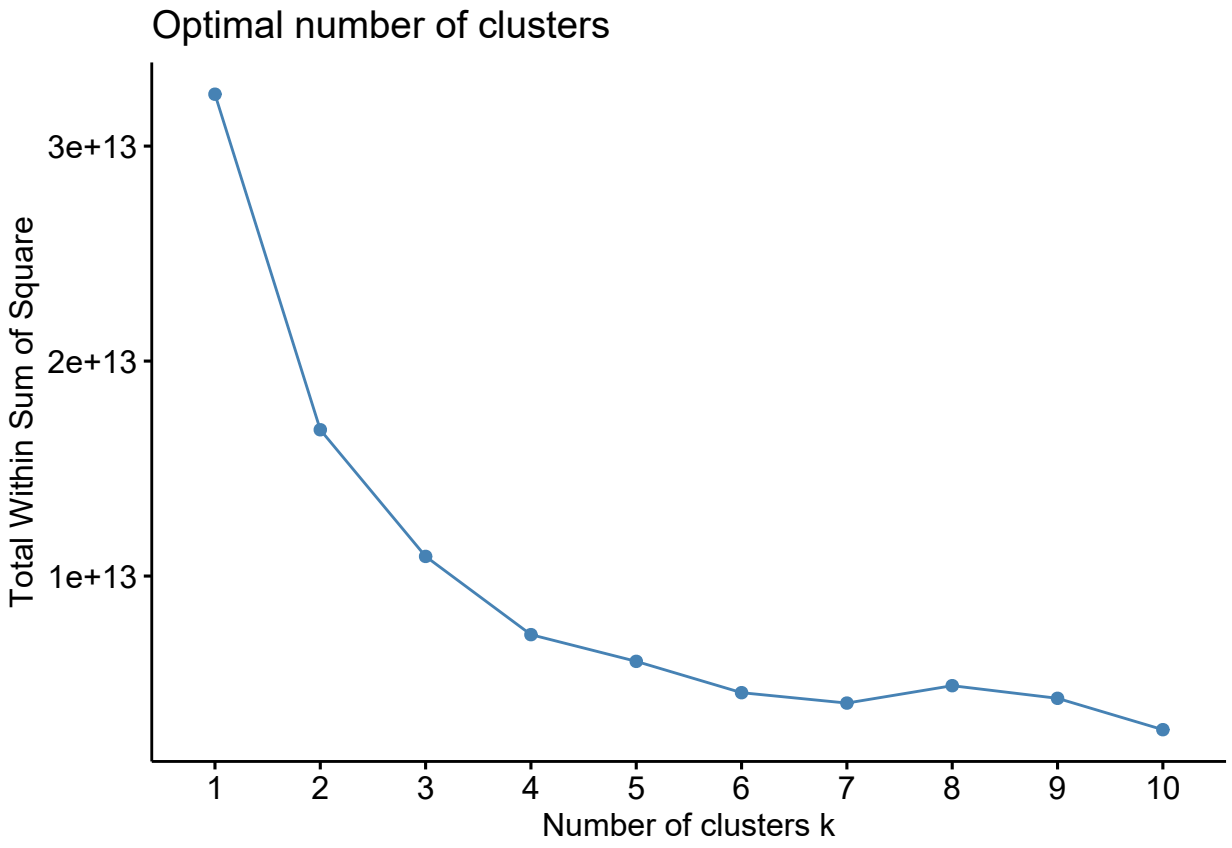


¿Cuál cree usted que se acomoda mejor a los datos?

1.5. Determinación del número de clústeres

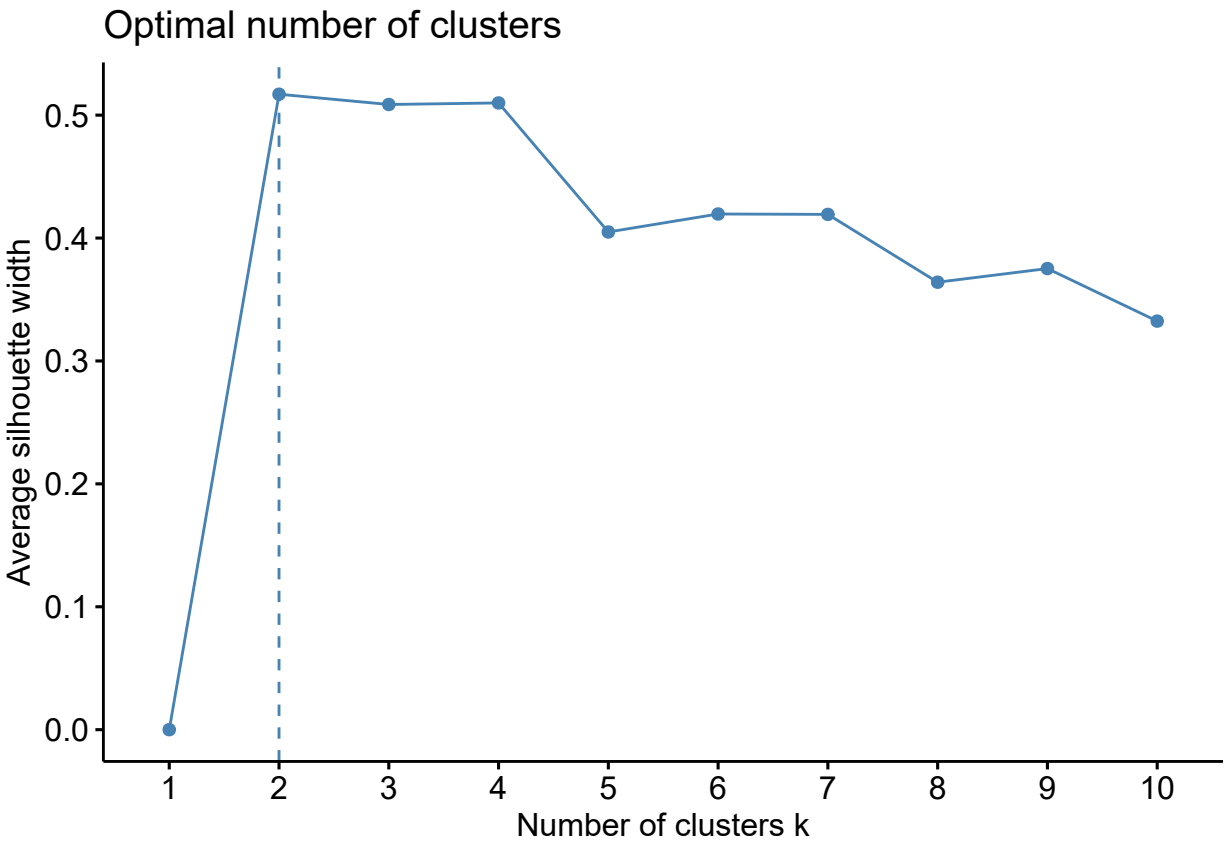
Anteriormente, determinamos el número de clústeres utilizando números predefinidos. Veamos ahora algunos algoritmos que nos pueden ayudar con esto:

```
set.seed(123)
fviz_nbclust(data_1, kmeans, method = "wss")
```



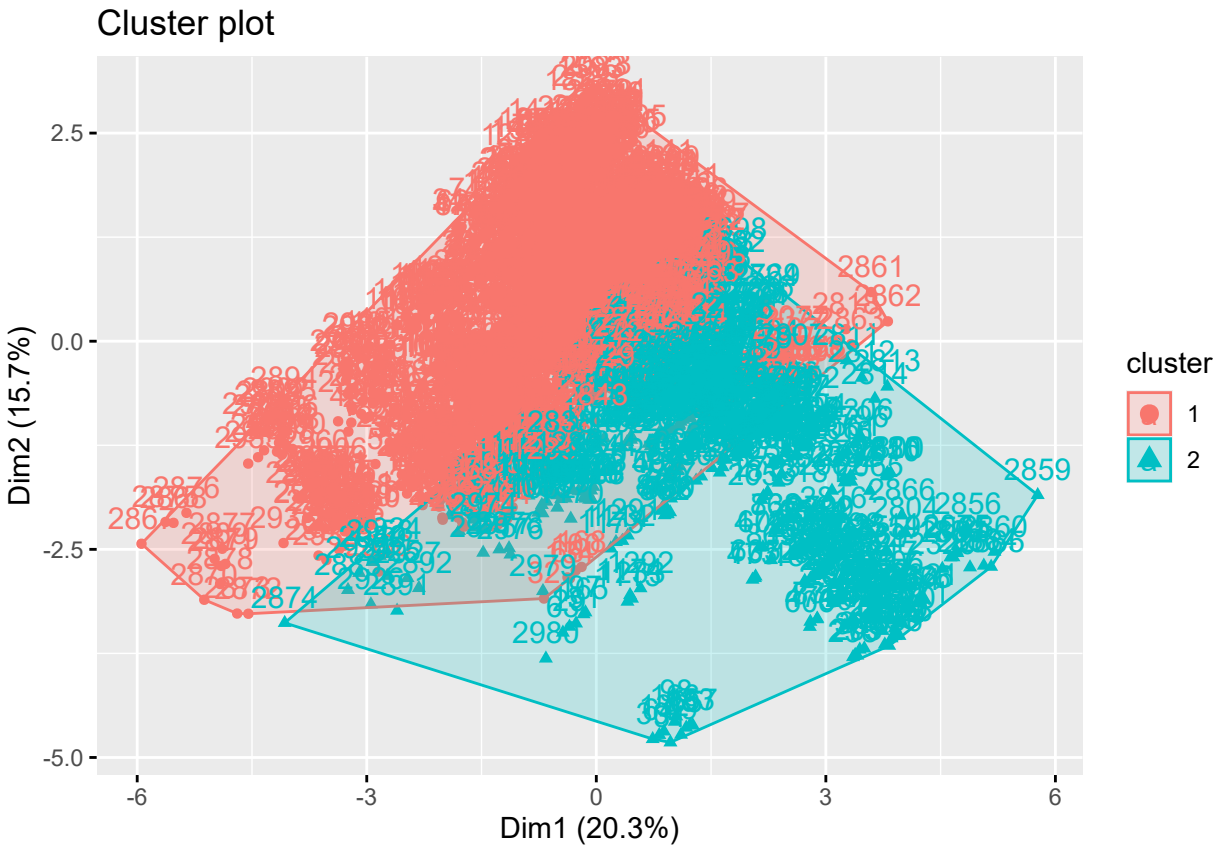
Un método distinto:

```
set.seed(123)
fviz_nbclust(data_1, kmeans, method = "silhouette")
```



Al parecer, en ambos casos tenemos evidencia de que 2 es el número correcto. Calculemos dicho caso:

```
set.seed(123)
final <- kmeans(data_1, 2, nstart = 25)
fviz_cluster(final, data = data_1)
```



2. Análisis de clústeres: Una aplicación en Power Bi

Veamos ahora si se puede replicar algo de lo realizado en el punto anterior dentro de Power Bi. Utilizaremos el mismo set de datos anterior, y nos guiaremos en base al siguiente ejemplo:

Enlace acá

3. Cambios en la Planificación diáctica

- Se elimina evaluación formativa
- Evaluación de participación se une con evaluación grupal. Dicha evaluación será para el 14 de junio.
- Sumativa II, queda para la misma semana y será individual. Será en base a preguntas de desarrollo, selección múltiple, verdadero o falso, y análisis de casos.
- Se debe comenzar a pensar en el posible set de datos a utilizar. Algunas opciones que hay son:

1. Futbol.
2. Baloncesto.
3. Kpop.
4. League of legends.
5. Análisis de clientes.

En base a dicho set de datos, deberán realizar una propuesta de análisis de datos, lo que corresponderá al 5 % de la evaluación de participación de la Unidad III. El procentaje restante, corresponderá a la presentación de su reporte. Todo esto se realizará de manera grupal.

4. Fechas Relevantes

Unidad	Evaluación	Ponderación	Fecha
Unidad I	~~Evaluación diagnóstica~~		25/03/2021
	~~Evaluación Individual Participación~~	(5 %)	05/04/2021
	~~Evaluación Grupal~~	(15 %)	27/04/2021 - 04/05/2021
	~~Evaluación Individual - Sumativa I~~	(30 %)	11/05/2021
Unidad II	~~Evaluación Formativa~~		13/05/2021
	~~Evaluación Individual Participación~~	(5 %)	27/05/2021
	Evaluación Grupal	(15 %)	14/06/2021
	Evaluación Individual - Sumativa II	(15 %)	17/06/2021
Unidad III	Evaluación Formativa		22/06/2021
	Evaluación Individual Participación	(5 %)	24/06/2021
	Evaluación Individual Sesión I- Sumativa III	(15 %)	08/07/2021
	Evaluación Individual Sesión II- Sumativa III	(15 %)	13/07/2021