DIPLOMADO EN CIENCIA DE DATOS

Módulo: Minería de Datos

Validación de clusterización

Universidad Nacional de Colombia

Contenido

- Validación de clusters
  - Validación interna
  - Validación externa
  - Validación relativa
- Validación de resultados

Validación de clusters

El agrupamiento es un método de análisis **no supervisado** porque no existe una manera de controlar/revisar qué tan precisos son los resultados

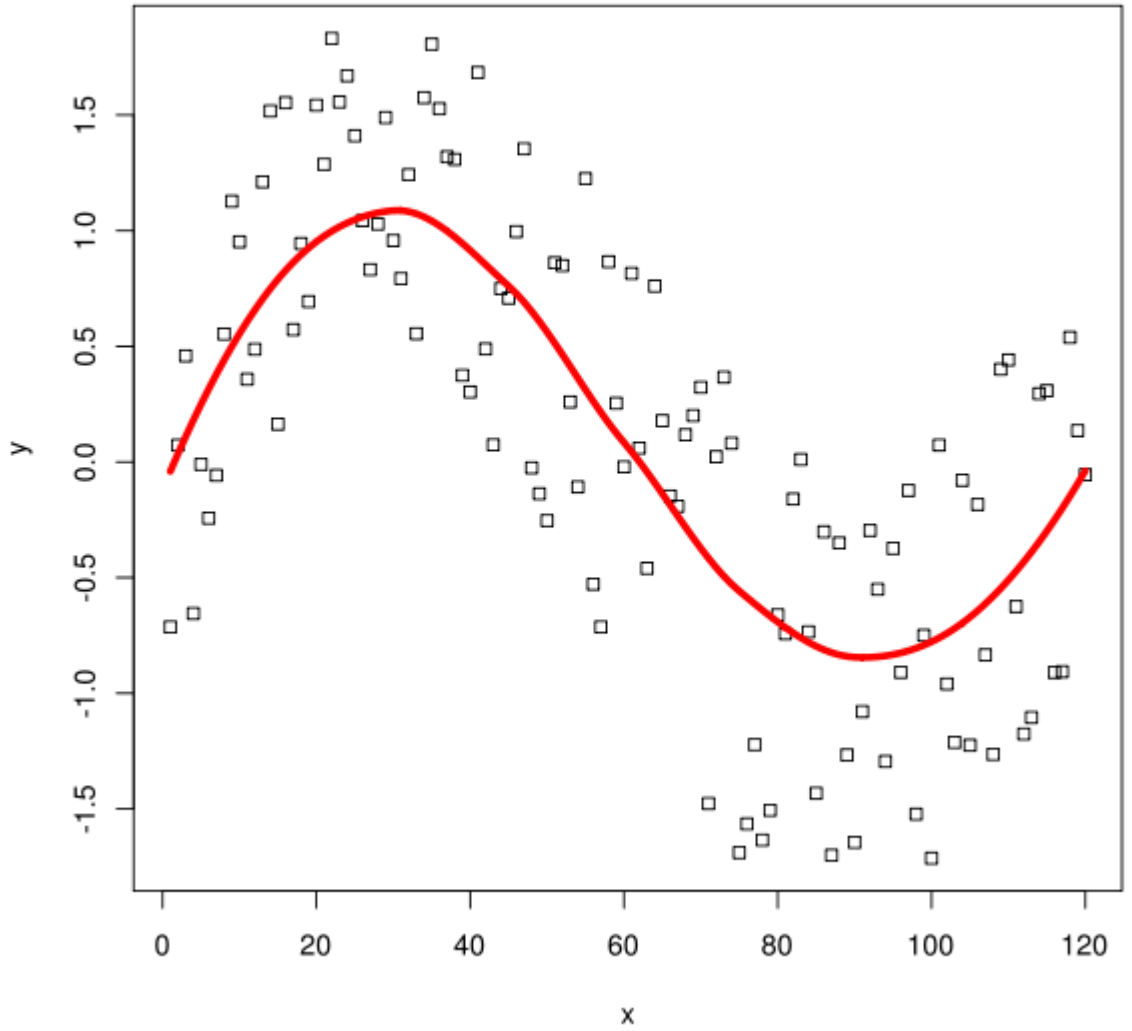
¿Cómo saber si la solución o los clusters definidos son adecuados?

La etiqueta de clusters es una variable ***inventada***

id	Edad	Género	Ingreso	Educación	Cluster
123	18	F	Bajo	1	2
456	25	O	Medio	2	2
789	26	M	Alto	3	1

Validación de clusters

El agrupamiento es un método de análisis **no supervisado** porque no existe una manera de controlar/revisar qué tan precisos son los resultados



Validación de clusters

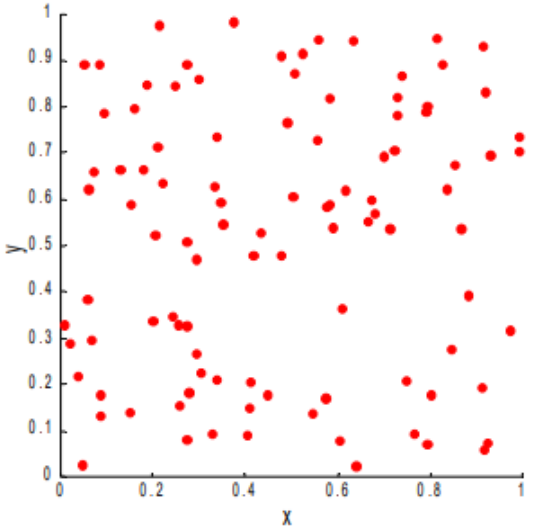
La validación de los clusters permite evaluar qué tan “buenos” son los clusters obtenidos.

- Evita encontrar patrones donde no los hay (datos aleatorios)
- Comparación de diferentes algoritmos de clusterización
- Comparación de soluciones de clusterización
- Escoger parámetros óptimos

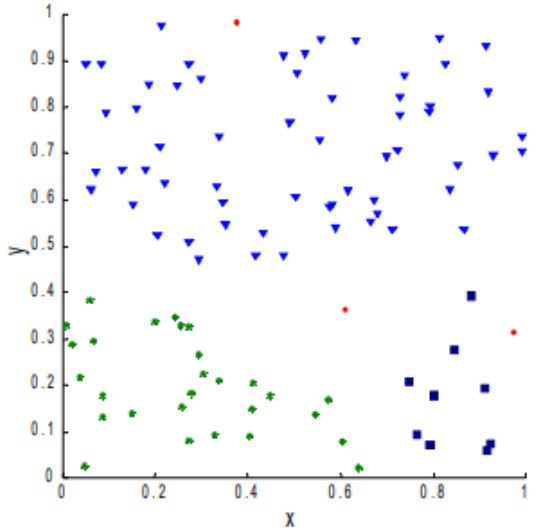
Tenemos tres tipos de validación: interna, externa y relativa

Validación de clusters

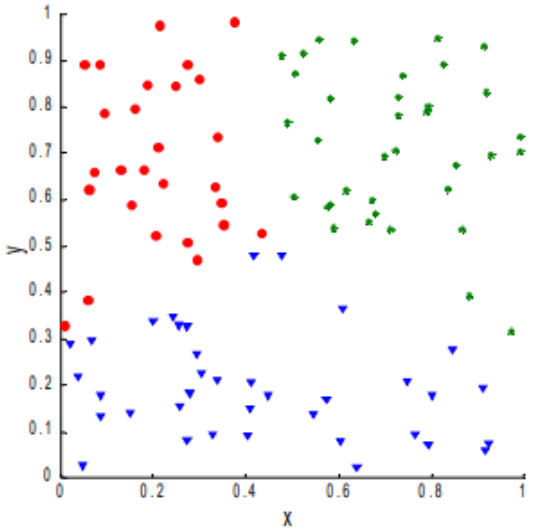
Random Points



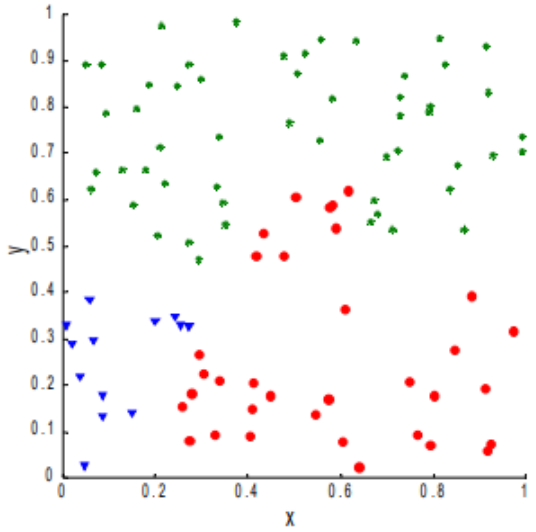
DBSCAN



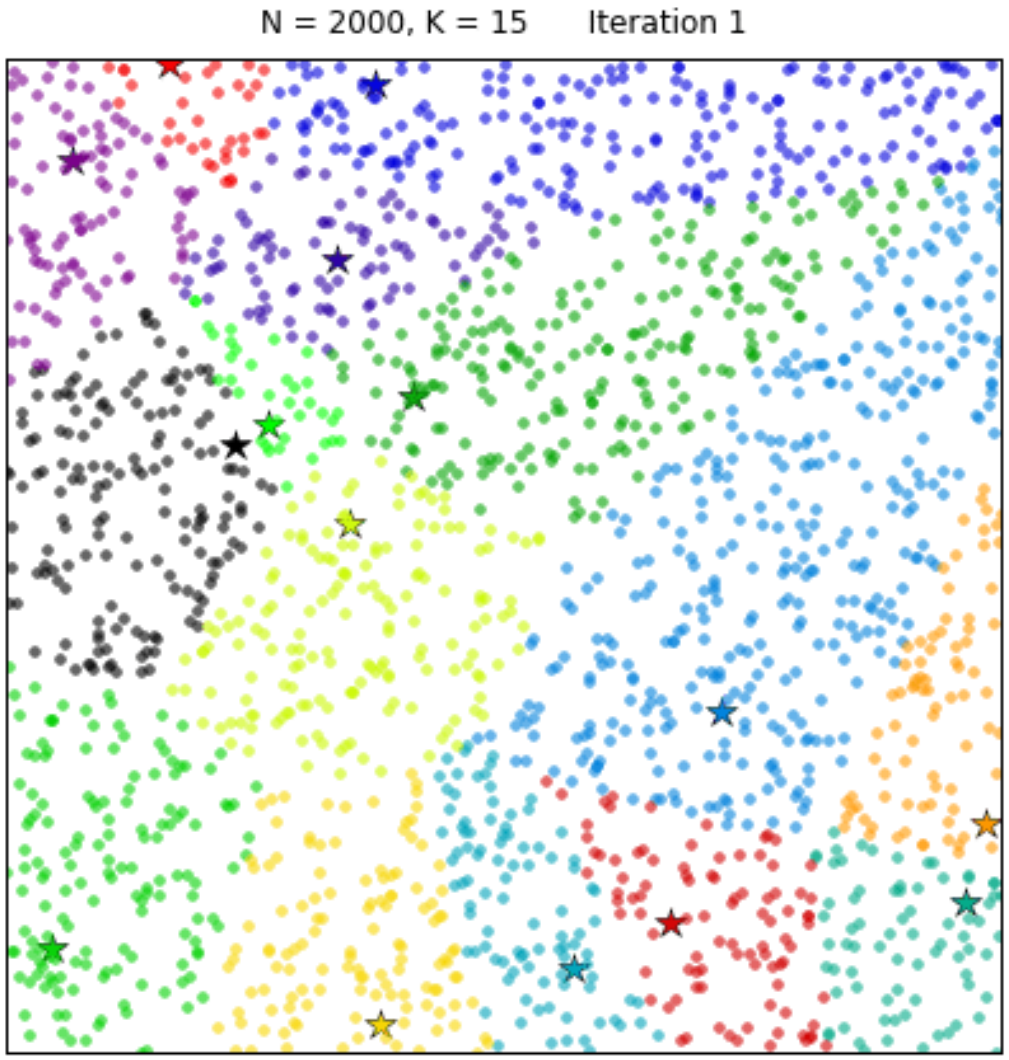
K-means



Complete Link



Validación de clusters



Validación de clusters

1.Validación interna: Utiliza la información interna del proceso de agrupamiento para evaluar qué tan buena es una estructura de agrupamiento sin referencia a información externa. Útil para escoger número de clusters o algoritmo de agrupación.

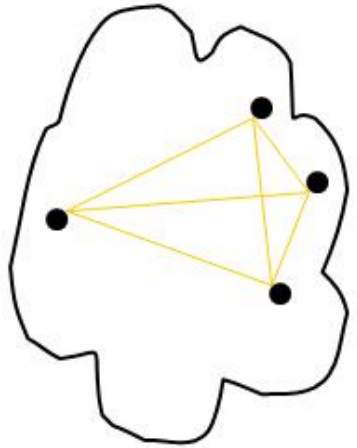
2.Validación externa: Compara las clases de la clusterización con un resultado conocido externo (label). Evalúa qué tanto coinciden los clusters con las clases conocidas. Se utiliza cuando ya se conoce categorías de grupos y se quiere seleccionar el algoritmo apropiado. Muy poco común.

3.Validación relativa: Evalúa la estructura de la clusterización cambiando los valores de algún parámetro en el mismo algoritmo. Determinar número óptimo de clusters

Validación Interna

Evalúa la cohesión, separación y conectividad

Cohesión: Mide qué tan cercanos son las observaciones dentro de un mismo clúster. Entre más cercanos sean mejor es la definición del cluster

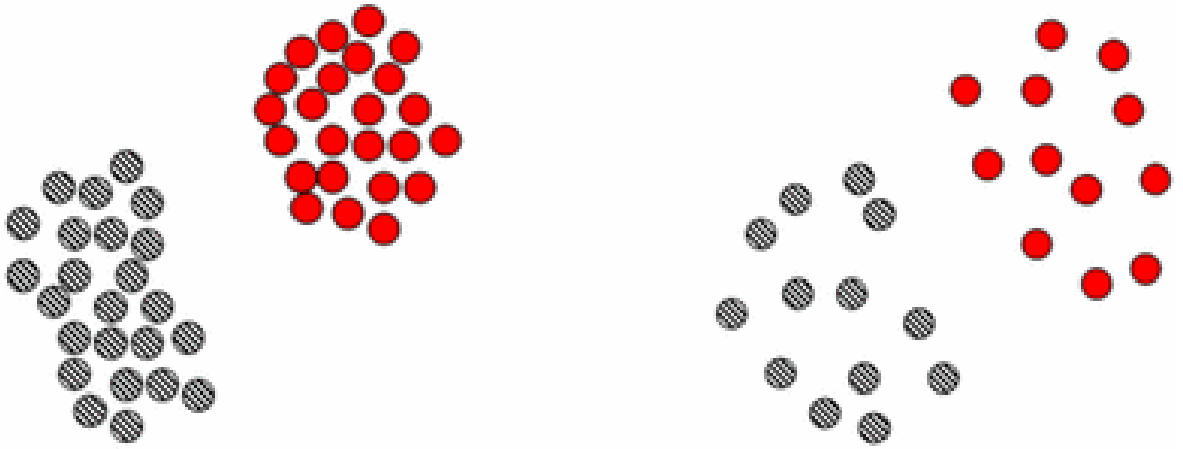


cohesion

Validación Interna

Evalúa la cohesión, separación y conectividad

Cohesión: Mide qué tan cercanos son las observaciones dentro de un mismo clúster. Entre más cercanos sean mejor es la definición del cluster



Validación Interna - Evaluación de cohesión

Sum of squares Within (SSW) – Dentro

Medida de validación interna, dónde verificamos la cohesión de los clusters

$$ssw = \sum_{i=1}^{k} \sum_{x \epsilon c_i} dist^2 (m_i, x)$$

Donde k es el número de clusters, x un punto dentro del cluster c_i y m_i corresponde al centroide del cluster c_i

Se busca que la suma de distancias entre los puntos de un mismo cluster sea mínima y un valor bajo de SSW indicaría una buena solución.

Validación Interna - Evaluación de cohesión

Sum of squares Within (SSW) – Dentro

Es una buena técnica para comparar dos agrupaciones o dos clusters

Pero ¿cómo cuantificar qué tan buena resultan las diferencias?

Puede ser utilizado también para definir el número de clusters (elbow method)

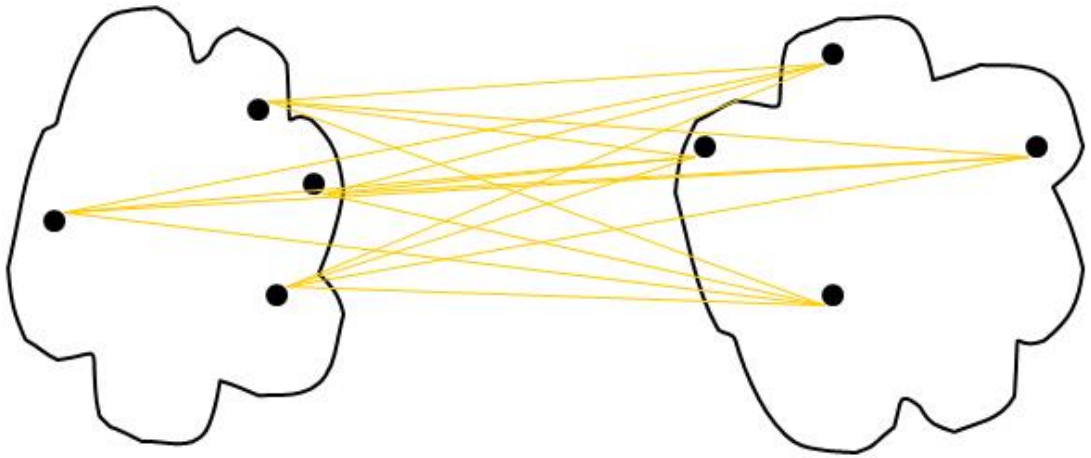
Falta de contexto para definir que es un “buen valor”. Un SSW de 20 es bueno, regular o malo?

Se puede comparar los valores de SSW de la data observada con data aleatoria generada y comparar

Validación Interna

Evalúa la cohesión, separación y conectividad

Separación: Mide qué tan bien separados están los clusters, unos de otros. Se puede evaluar de diferentes maneras: distancia entre centroides, distancia mínima entre clusters, distancia máxima entre clusters

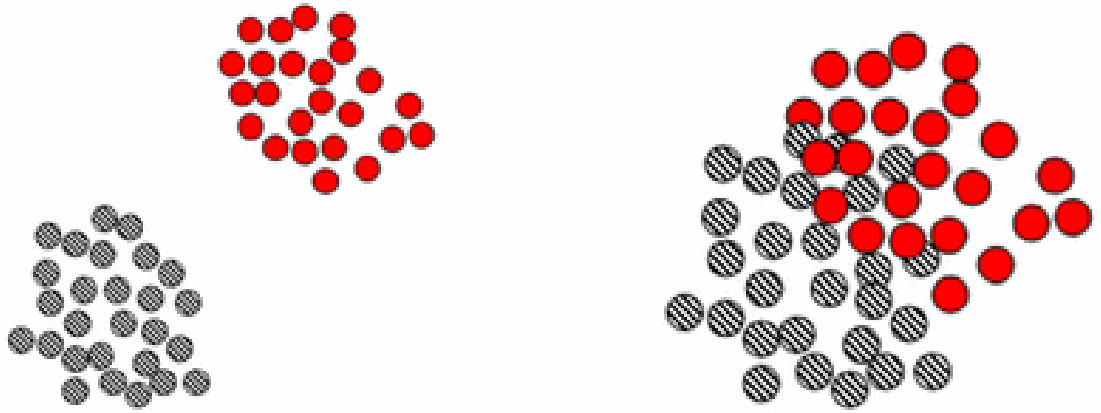


separation

Validación Interna

Evalúa la cohesión, separación y conectividad

Separación: Mide qué tan bien separados están los clusters, unos de otros. Se puede evaluar de diferentes maneras: distancia entre centroides, distancia mínima entre clusters, distancia máxima entre clusters



Validación Interna - Evaluación de separación

Sum of squares Between (SSB) – Entre

Medida para validación interna, dónde verificamos la separación de los clusters

$$SSB = \sum_{i=1}^{k} n_i dist^2(m_i - \bar{x})$$

Donde k es el número de clusters, n_i es el tamaño del cluster i, m_i corresponde al centroide del cluster i y $\bar{x}$ es la media

Se busca que la distancia entre los clusters sea máxima y un valor alto de SSB indicaría una buena solución.

Validación Interna - Otros índices

Coeficiente de Silhouette

Este coeficiente considera tanto la cohesión como la separación, para los puntos individuales y los clusters. Mide la calidad del agrupamiento

Para cada punto i:
Calcular a: distancia media del punto i a todos los puntos dentro de su cluster
Calcular b: min(distancia media del punto i a todos los demás puntos en otros clusters)

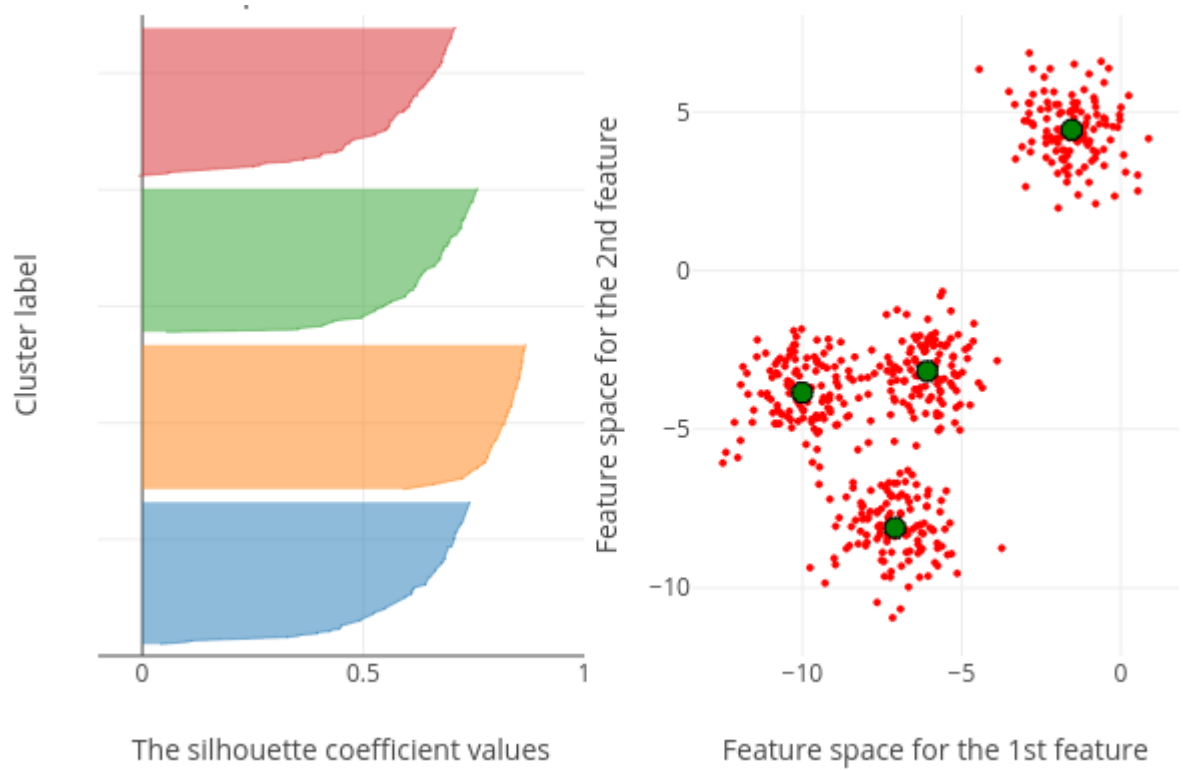
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Determina qué tan bien clasificado está un punto en su cluster
S(i) toma valores entre -1 y 1.
Siendo -1 una mala solución y 1 una buena solución.

Validación Interna - Otros índices

Coeficiente de Silhouette

Medimos qué tan bien asignado esta un punto en un cluster.



S(i) toma valores entre -1 y 1.
Siendo -1 una mala solución y 1 una buena solución.

Validación Interna - Otros índices

Coeficiente de Silhouette

El coeficiente para todos los clusters es

$$SC = \frac{1}{N}\sum_{i=1}^{N} s(i)$$

S(i) toma valores entre -1 y 1.
Siendo -1 una mala solución y 1 una buena solución.

Validación Interna

Evaluamos la estructura interna del agrupamiento para saber qué tan bueno es, teniendo en cuenta que queremos clusters:

- Homogéneos dentro (cohesión)
- Heterogéneos entre (separación)

Podemos hacerlo revisando suma de distancias dentro de clusters (cohesión), suma de distancia entre clusters (separación), coeficiente de silhouette (cohesión y separación)

Validación relativa

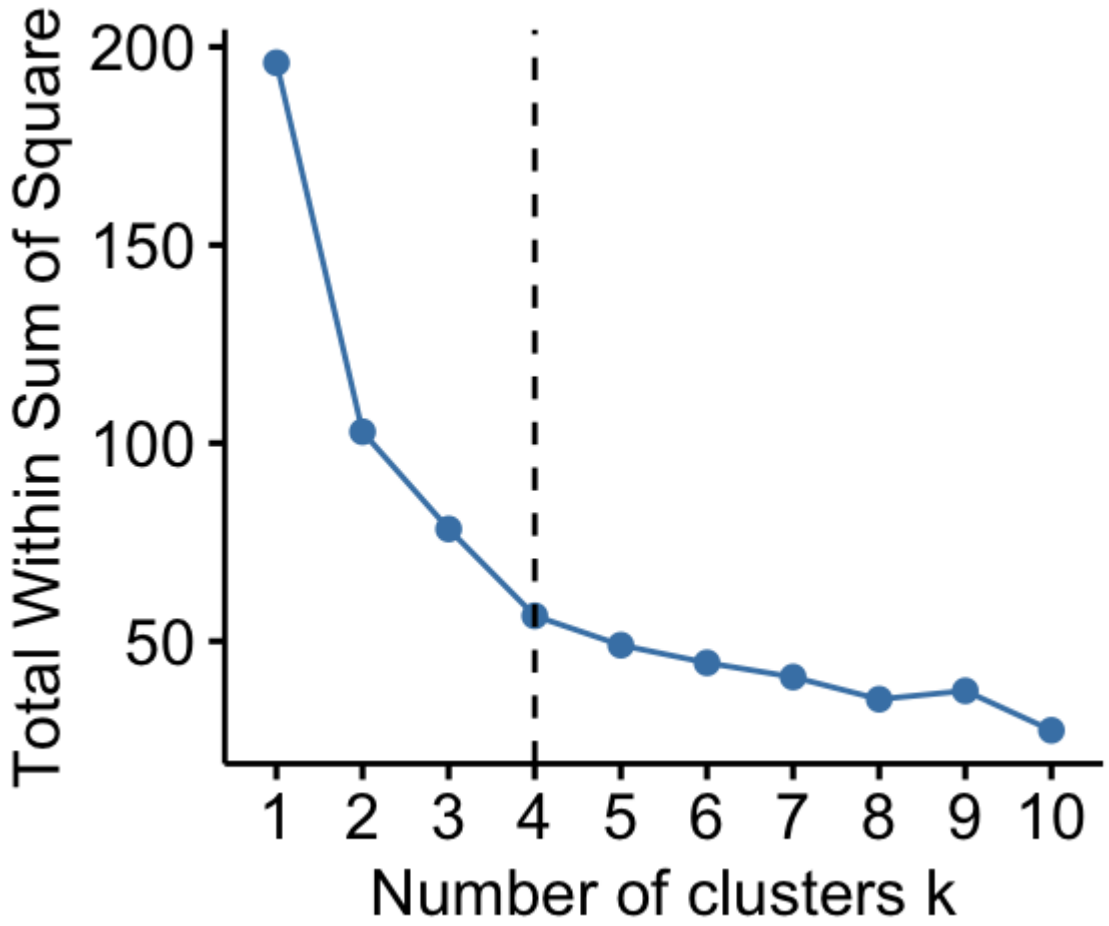
Evaluar la estructura del agrupamiento al probar diferentes parámetros para el mismo algoritmo

Se pueden utilizar medidas de validación interna o externa para comparar soluciones

Se usa para determinar valores óptimos de los parámetros.

Validación relativa

Evalúa la estructura de la clusterización cambiando los valores de algún parámetro en el mismo algoritmo.



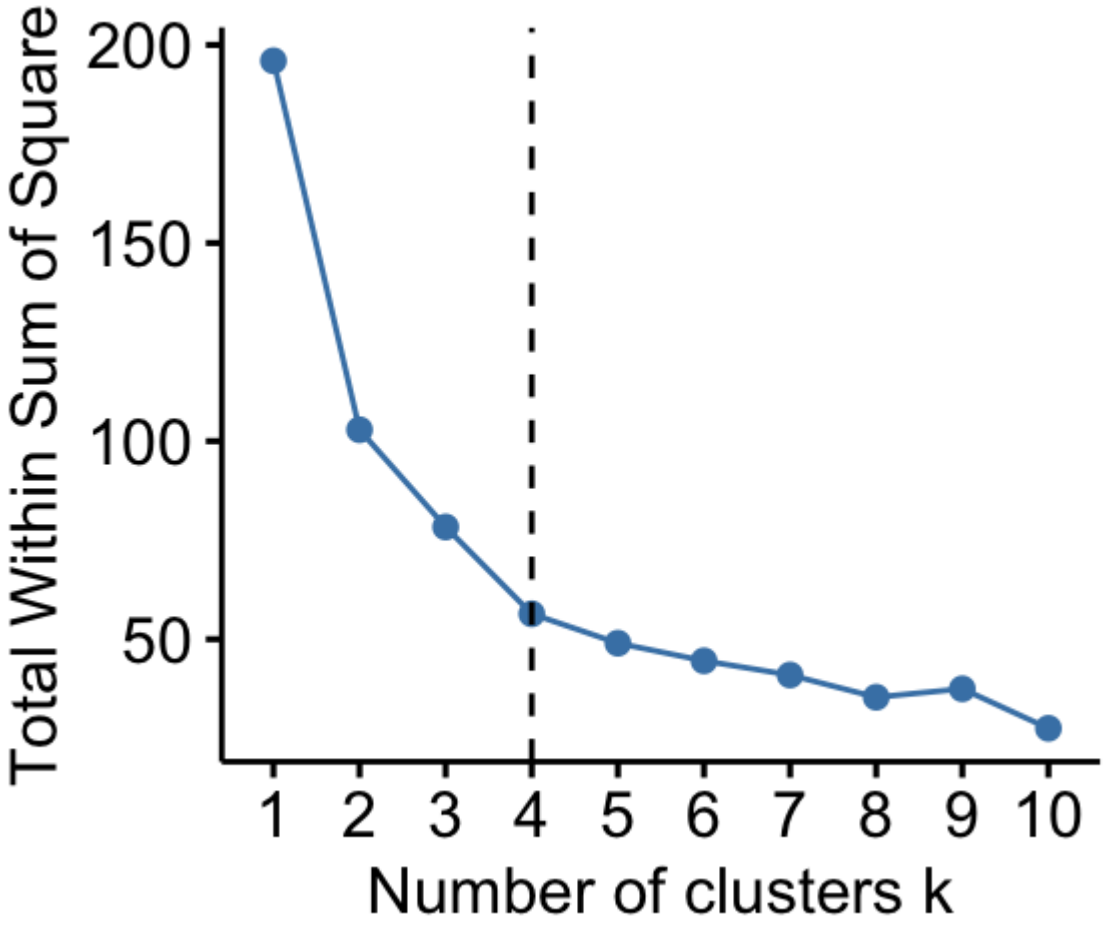
Validación relativa
Elección de número de clusters – Método del codo (Elbow method)

1.Ajustar k-means para diferentes valores de k. Por ejemplo, k de 1 a 10 clusters

2.Para cada k, calcular el SSW

3.Graficar la curva del SSW para cada valor de k.

4.El punto en la curva donde se genere un ‘codo’ es considerado el número más apropiado para k

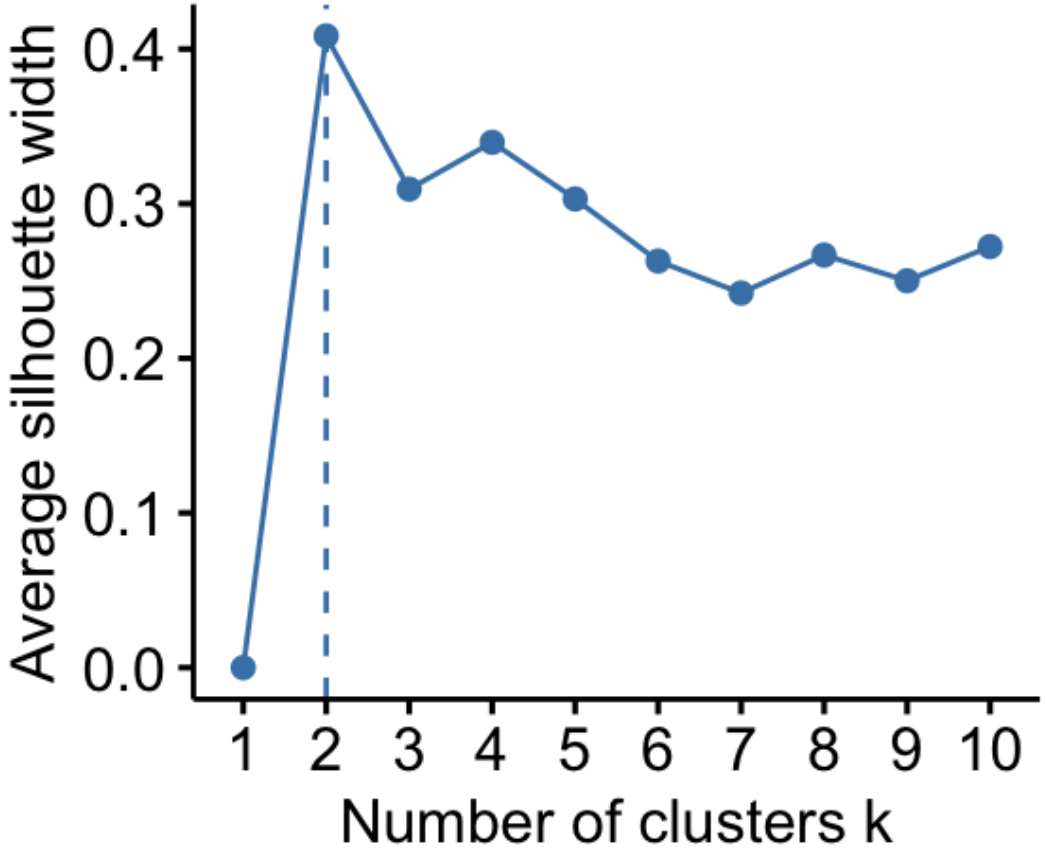


Validación relativa

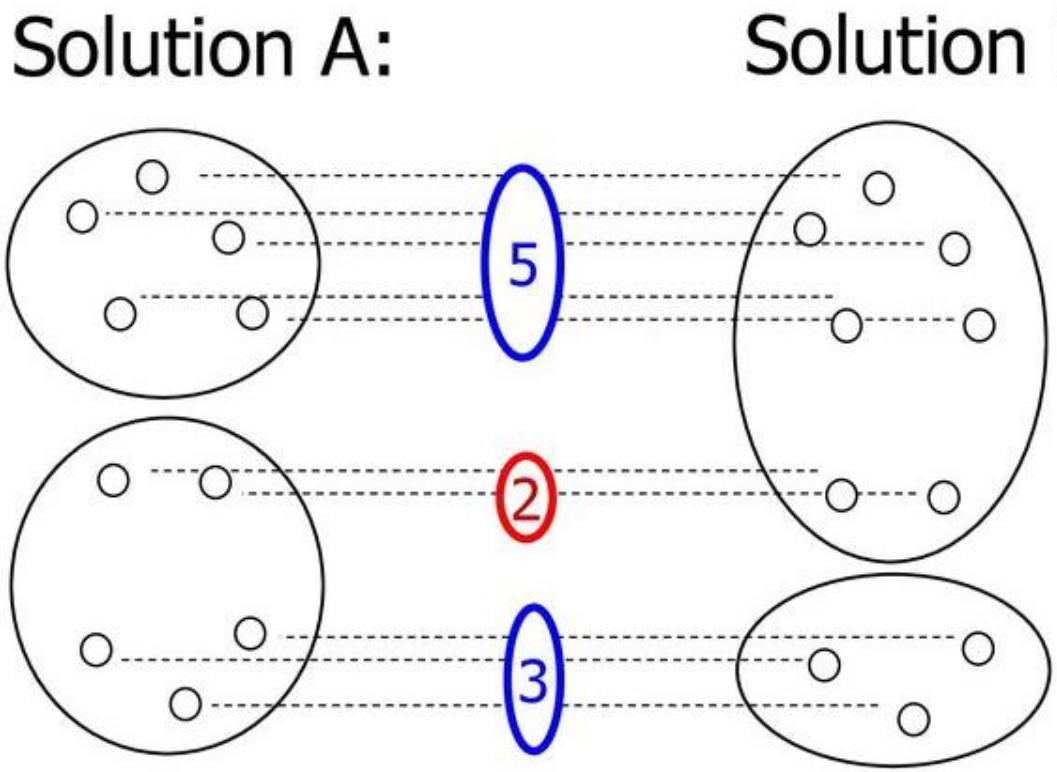
Coeficiente de Silhouette

También es un método utilizado para decidir el número de clusters, calculando el coeficiente de silhouette promedio y graficándolo para diferentes valores de k (cómo en el método del codo)

Aquí buscamos el valor de k para el cual el promedio del coeficiente de silhouette es máximo



Validación externa



Apriori se conoce la clase de cada observación y se puede contrastar con la solución

No es usual conocer esta clasificación, si se tiene se emplearían otros métodos de análisis supervisado.

Validación externa

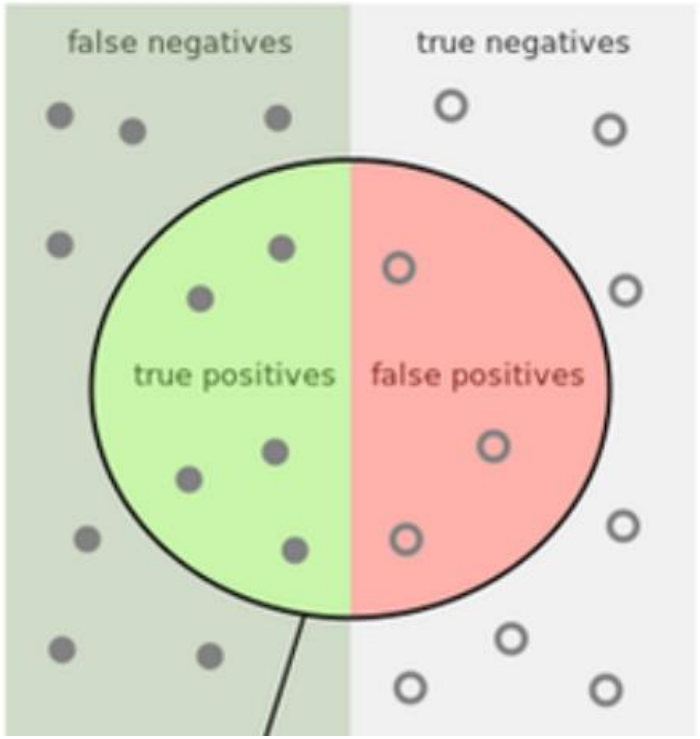
Matriz de confusión

		Clase verdadera	
		C1	C2
Clase Clustering	C1	Verdadero Positivo	Falso Positivo
	C2	Falso Negativo	Verdadero Negativo

Validación externa

Matriz de confusión

		Clase verdadera	
		C1	C2
Clase Clustering	C1	a	b
	C2	c	d



$$Precisión = \frac{a}{a+b}$$

$$Precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$Recall = \frac{a}{a+c}$$

$$Recall = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Validación externa

Entropía

La entropía de cada cluster esta dado por

$$entropía(D_i) = -\sum_{j=1}^{k} Pr_i(c_j) log_2 Pr_i(c_j)$$

Dónde $Pr_i(c_j)$ corresponde a la proporción de puntos de la clase j (c_j clases reales) que quedan ubicados en el cluster i (D_i clases creadas por la agrupación)

La entropía total es

$$entropía(D) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} * entropía(D_i)$$

Mide la 'pureza' de cada cluster. Valores cercanos a 0 indican una alta pureza

Validación de resultados

¿Tienen sentido los resultados obtenidos?

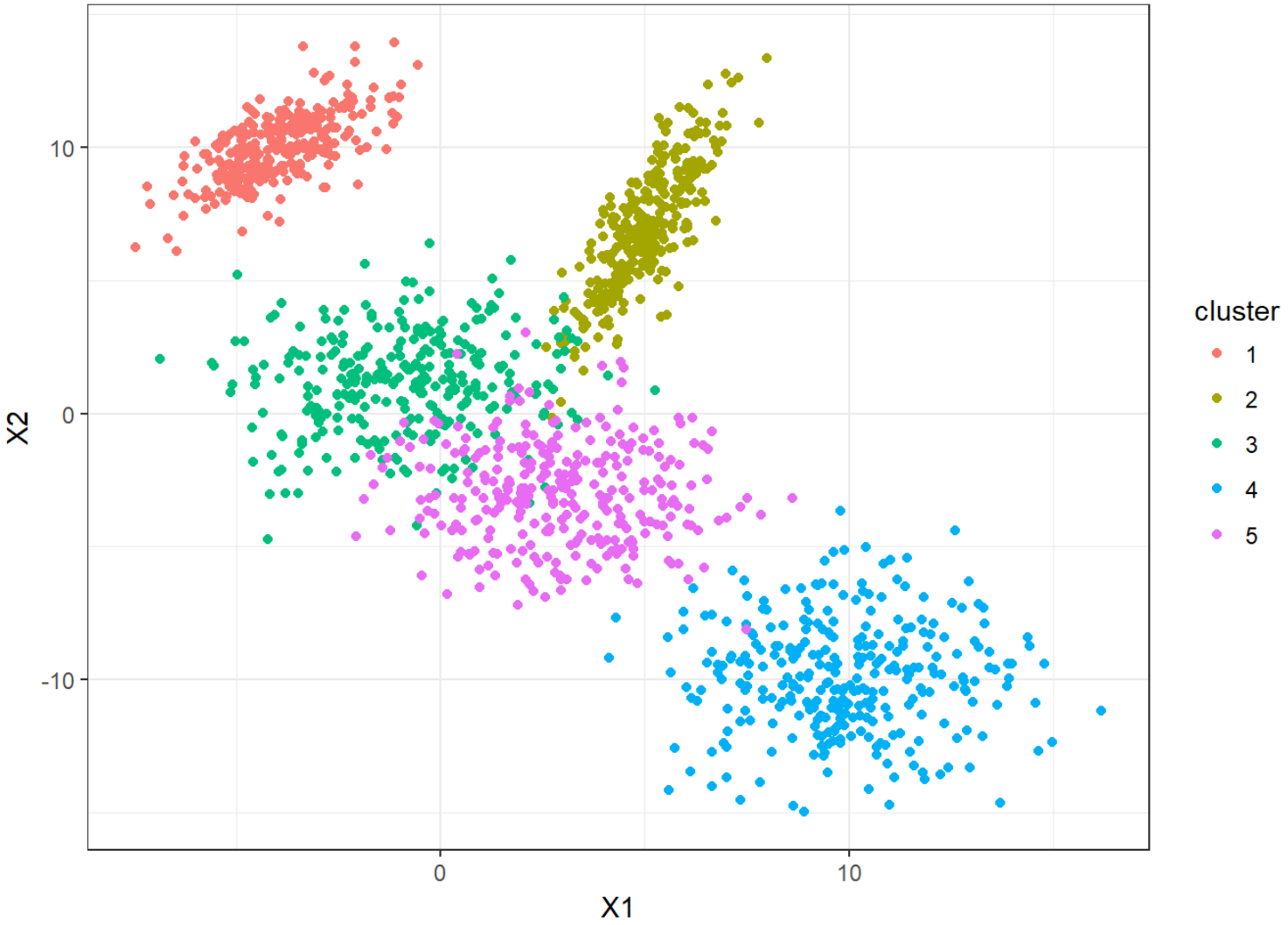
¿ Para qué utilizaremos los clusters?

¿ Qué hace diferentes a los clusters?

Validación de resultados

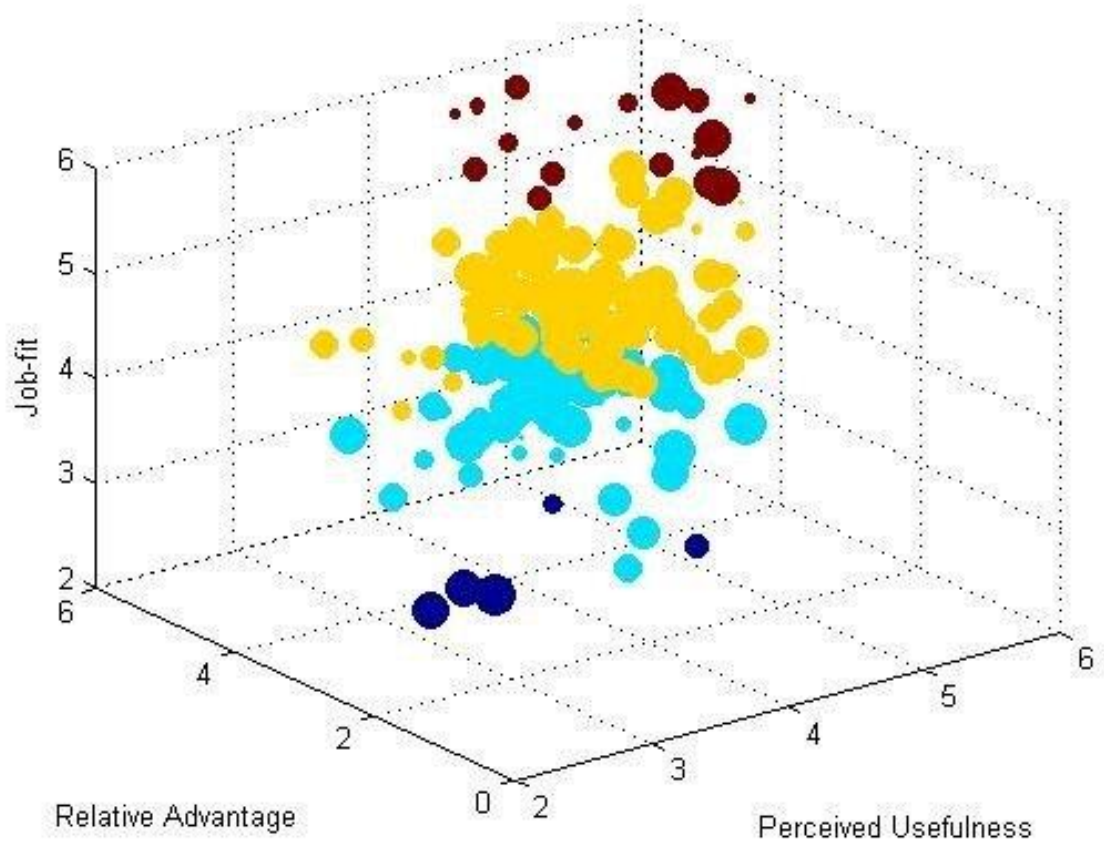
Tamaño de clusters

Cluster	Tamaño	Participación
1	100	11%
2	120	13%
3	200	21%
4	250	26%
5	280	29%



Validación de resultados

Visualizar los resultados con las variables, tiene sentido?



Validación de resultados

Estadísticas resumen

Variables	Cluster 1	Cluster 2	Overall
Q7: How many indivi...	4.20	2	3.38
Doctoral degree	40.0%	0.0%	25.0%
Less than high schoo...	0.0%	100.0%	37.5%
Master's degree	60.0%	0.0%	37.5%
Divorced	0.0%	66.7%	25.0%
Married	100.0%	0.0%	62.5%
Separated	0.0%	33.3%	12.5%
Q5: What is your age?	32.4	50.3	39.1

Combinaciones

- Si no tengo información de contexto hacer un agrupamiento jerárquico y definir el número de clusters de la información del dendograma, luego ajustar un kmeans con este k

- Si hay muchos datos, hacer una clusterización con una muestra aleatoria. Asignar el cluster a los puntos que no quedaron en la muestra utilizando métodos de clasificación

- Aplicar un ACP, tomar las coordenadas de los puntos en los componentes como variables para hacer un kmeans. https://365datascience.com/pca-k-means/

- Si los datos son categóricos, ajustar un Análisis de Correspondencias Múltiples – ACM (un ACP en datos categóricos), y utilizar las coordenadas de los puntos para hacer un kmeans. https://www.kaggle.com/rekahalmai/mca-clusters-and-k-means

Conclusiones

- Métodos para validar los resultados de la clusterización
- Escoger parámetros óptimos para la solución
- Usar conocimiento experto para validar resultados
- Combinaciones para hacer agrupamiento