

Estadística en Analítica

2023-2

Pablo A. Saldarriaga
psaldar2@eafit.edu.co

UNIVERSIDAD
EAFIT

¿Dudas del Taller 2?

Técnicas no supervisadas

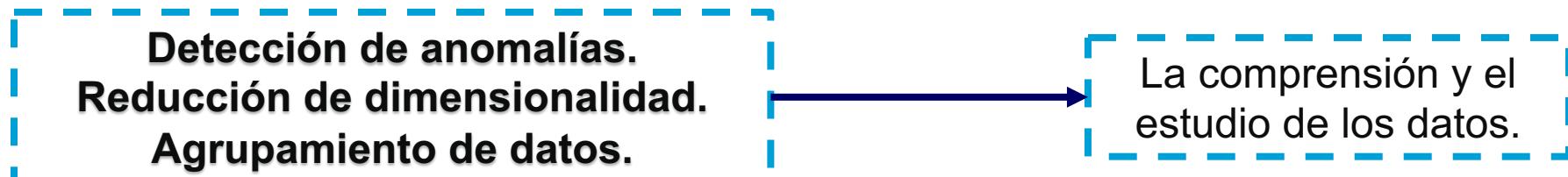
Aprendizaje no supervisado

El proceso de modelado se realiza sobre un **conjunto de datos** formado por solo **entradas** al sistema.

Sin conocimiento a priori de las observaciones.

El objetivo principal es descubrir aspectos interesantes sobre las mediciones que se hacen:

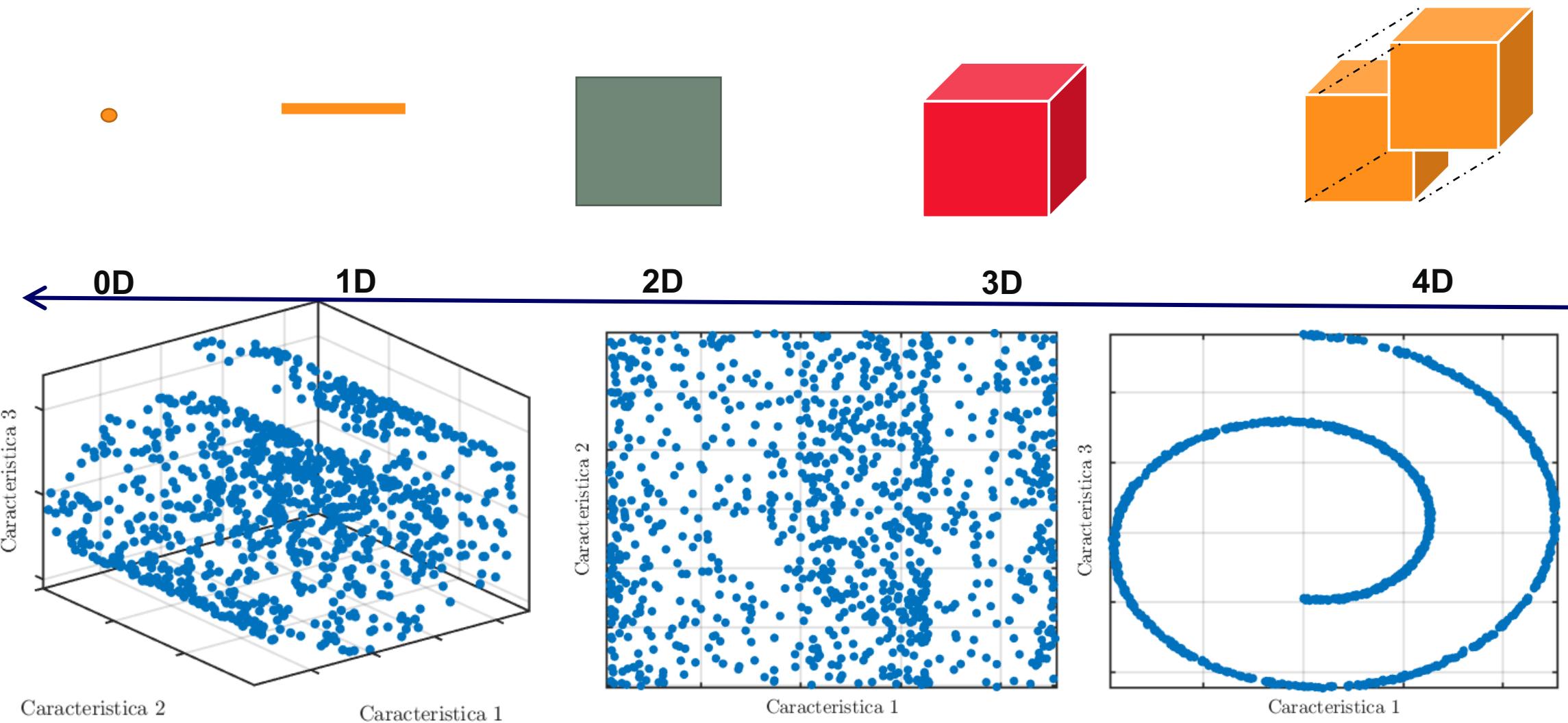
- ✓ ¿Hay alguna manera de visualizar la información que sea informativa?
- ✓ ¿Podemos descubrir subgrupos entre las variables o las observaciones?



Aprendizaje no supervisado: Retos

- ✓ El aprendizaje no supervisado es más subjetivo que el aprendizaje supervisado (ya que no hay un único objetivo para el análisis, tal como una variable respuesta)
- ✓ Ha crecido su aplicabilidad en áreas como:
 - ✓ Subgrupos de cáncer de seno en pacientes dependiente de la medición de los genes
 - ✓ Características de compradores por sus características de búsqueda o comportamiento de compra
 - ✓ Películas agrupadas por ratings y espectadores
- ✓ En ocasiones es más sencillo encontrar data sin etiquetas para ser analizadas

Reducción de dimensionalidad



Reducción de dimensionalidad

- ✓ **Eliminar** algunas características:

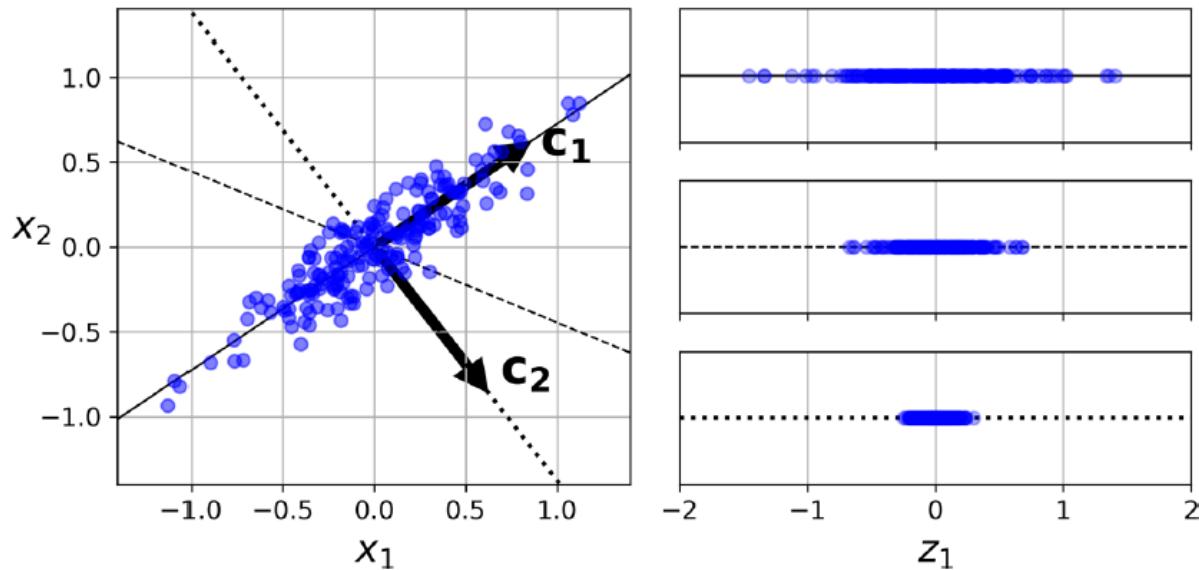
¿Cuáles son las menos importantes?

Podríamos estar perdiendo información importante.

- ✓ **Extracción/modificación de Características**

Si tenemos 10 características crearemos otras 10 características nuevas e independientes (**combinación de las características originales**).

Análisis de Componentes Principales



- ✓ Es un algoritmo de reducción de dimensionalidad (o de transformación de la información).
- ✓ Método basado en la matriz de varianzas y covarianzas
- ✓ Proyección de los datos en direcciones de mayor variabilidad

*Imagen tomada de Aurelien Geron, *Hands on ML* 2019.

Análisis de Componentes Principales

- ✓ El primer componente principal de un conjunto de datos es la combinación lineal normalizada de las variables:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p$$

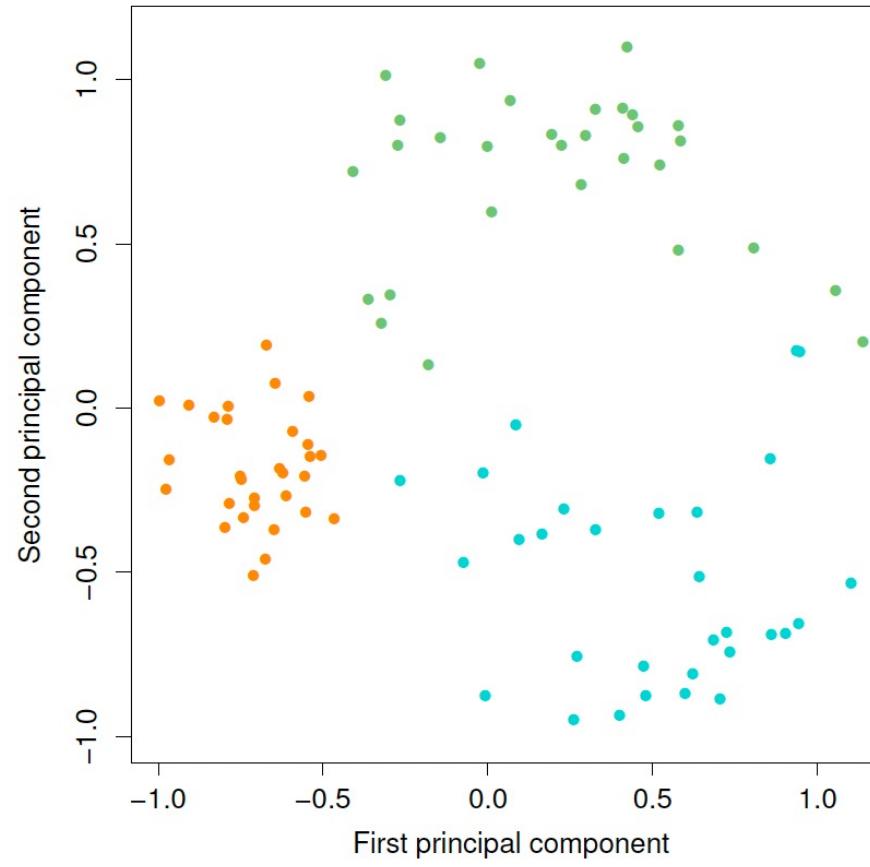
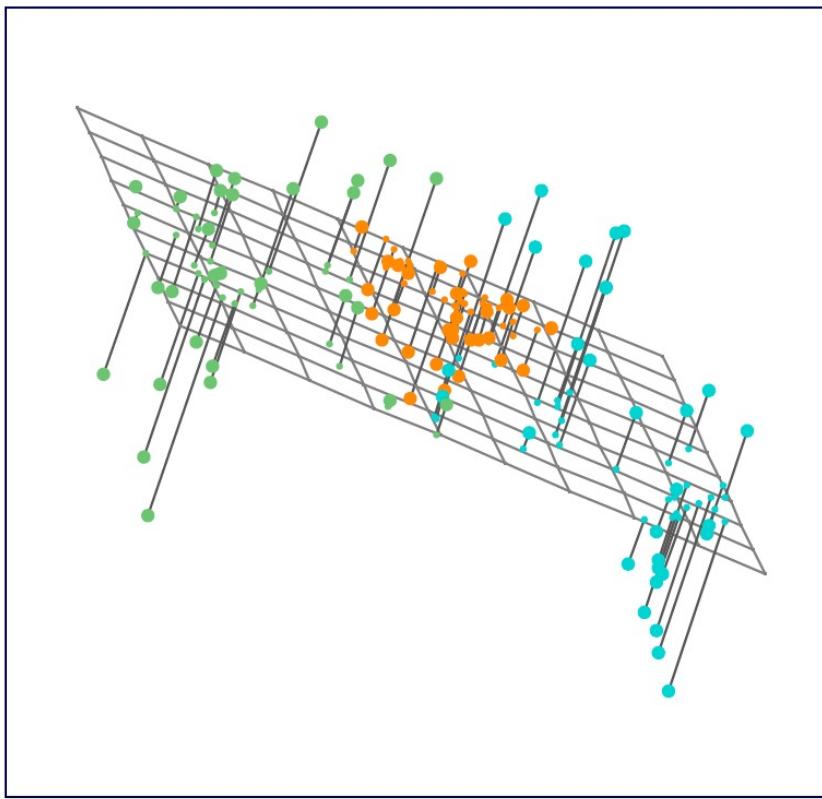
- ✓ Nos referimos a los elementos $\phi_{11}, \dots, \phi_{p1}$ como los pesos del primer componente principal, los cuales componen el vector asociado al primer componente principal. Estos definen la dirección en la cual los datos son proyectados
- ✓ Estos pesos son normalizados en norma 2

Análisis de Componentes Principales

- ✓ El segundo componente principal maximiza la variabilidad de la combinación lineal, considerando que es incorrelado con el primer componente principal

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \dots + \phi_{p2}x_{ip};$$

Análisis de Componentes Principales



Análisis de Componentes Principales

- ✓ La dirección del primer componente principal, tiene la propiedad de que define la linea que es más cercano a todas las observaciones (considerando la norma euclidea)
- ✓ La noción de los componentes principales se extiende a las p-dimensiones, en donde los k componentes tomados, corresponden al hiperplano p-dimensional más cercano a las observaciones

Análisis de Componentes Principales

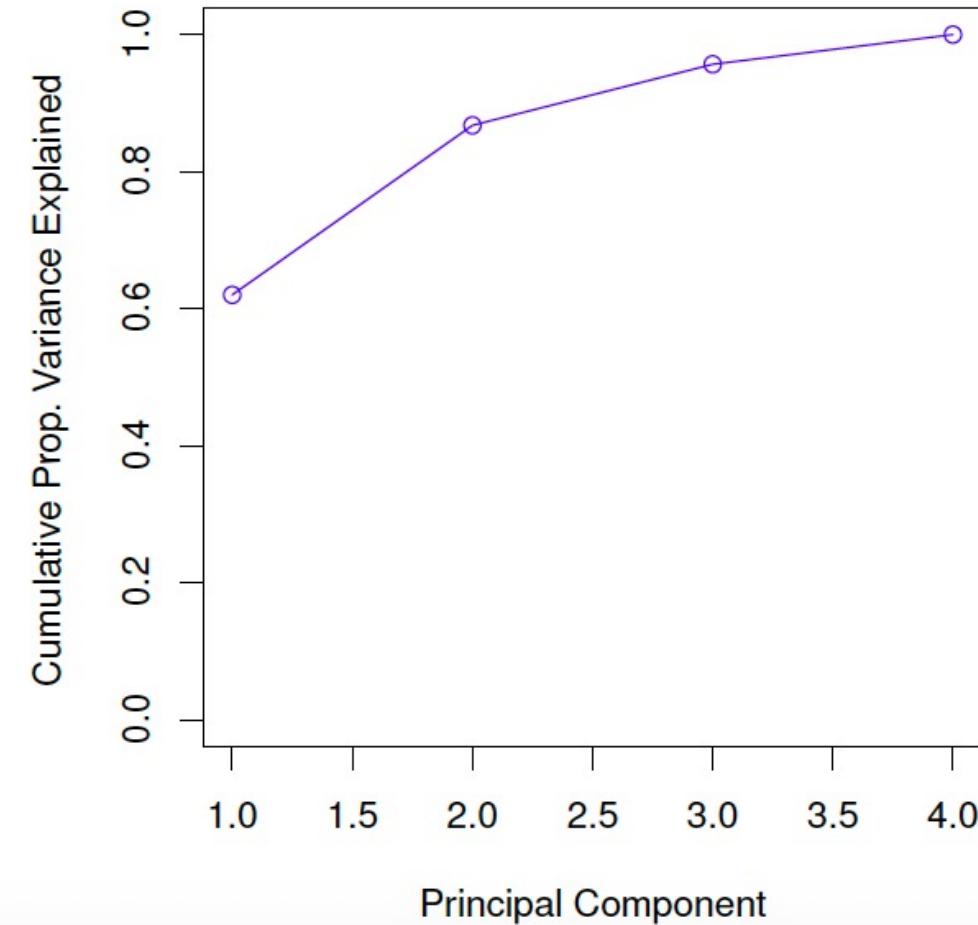
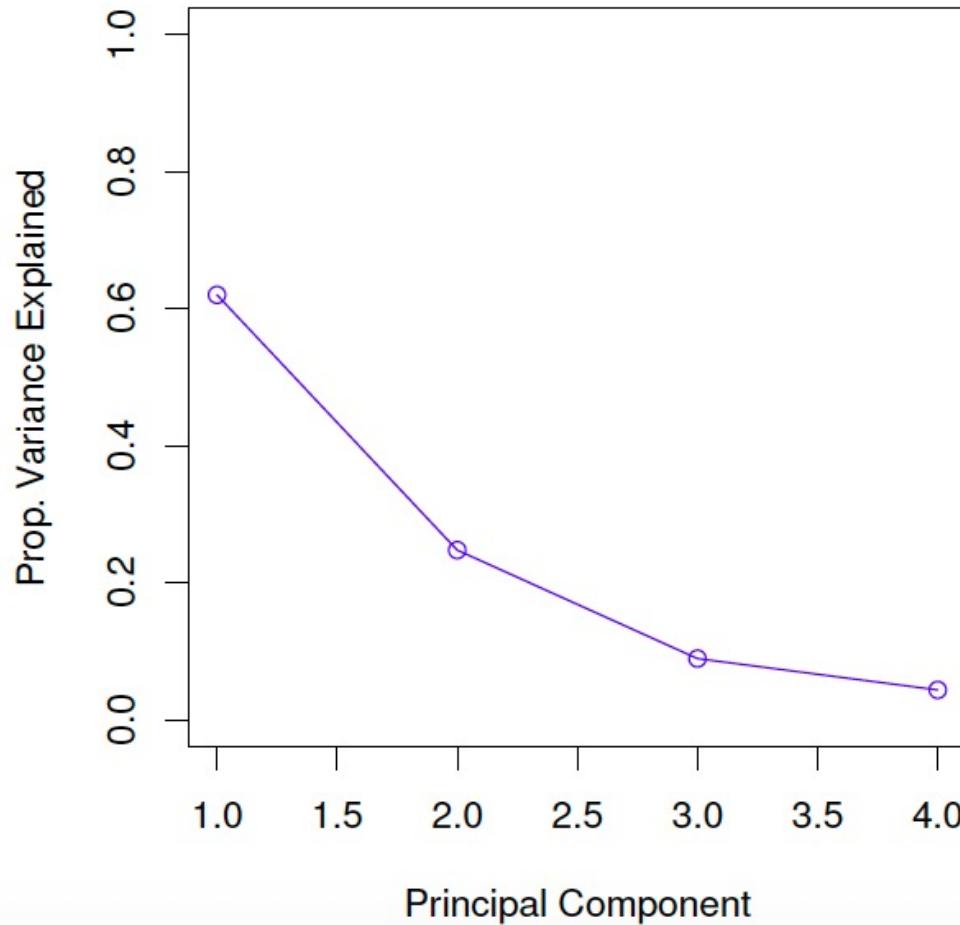
1. Estandarizar los datos de entrada.
2. Obtener los autovectores y autovalores de la matriz de covarianza (descomposición espectral)
3. Ordenar los autovalores de mayor a menor y elegir los “k” autovectores que se correspondan con los autovectores “k” más grandes.
4. Construir la matriz de proyección W con los “k” autovectores seleccionados.
5. Transformamos el dataset original X para obtener las nuevas características k-dimensionales.

Análisis de Componentes Principales

- ✓ Para entender la fuerza de cada componente, nos interesa saber la proporción de varianza explicada de cada componente
- ✓ La varianza total asume que el conjunto de datos se encuentra centrado, así podemos definirla como:

$$\sum_{j=1}^p \text{Var}(X_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2$$

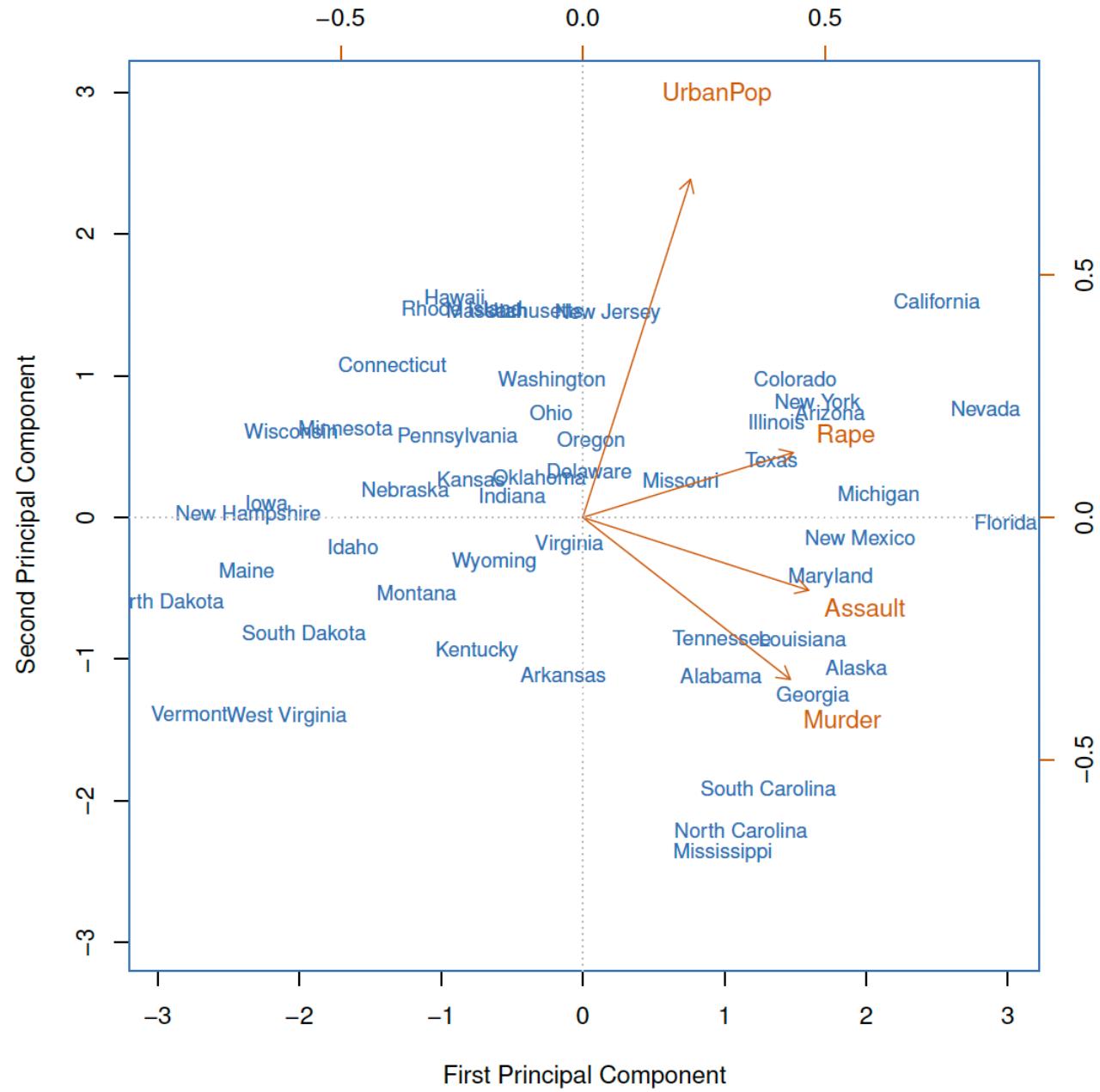
Análisis de Componentes Principales



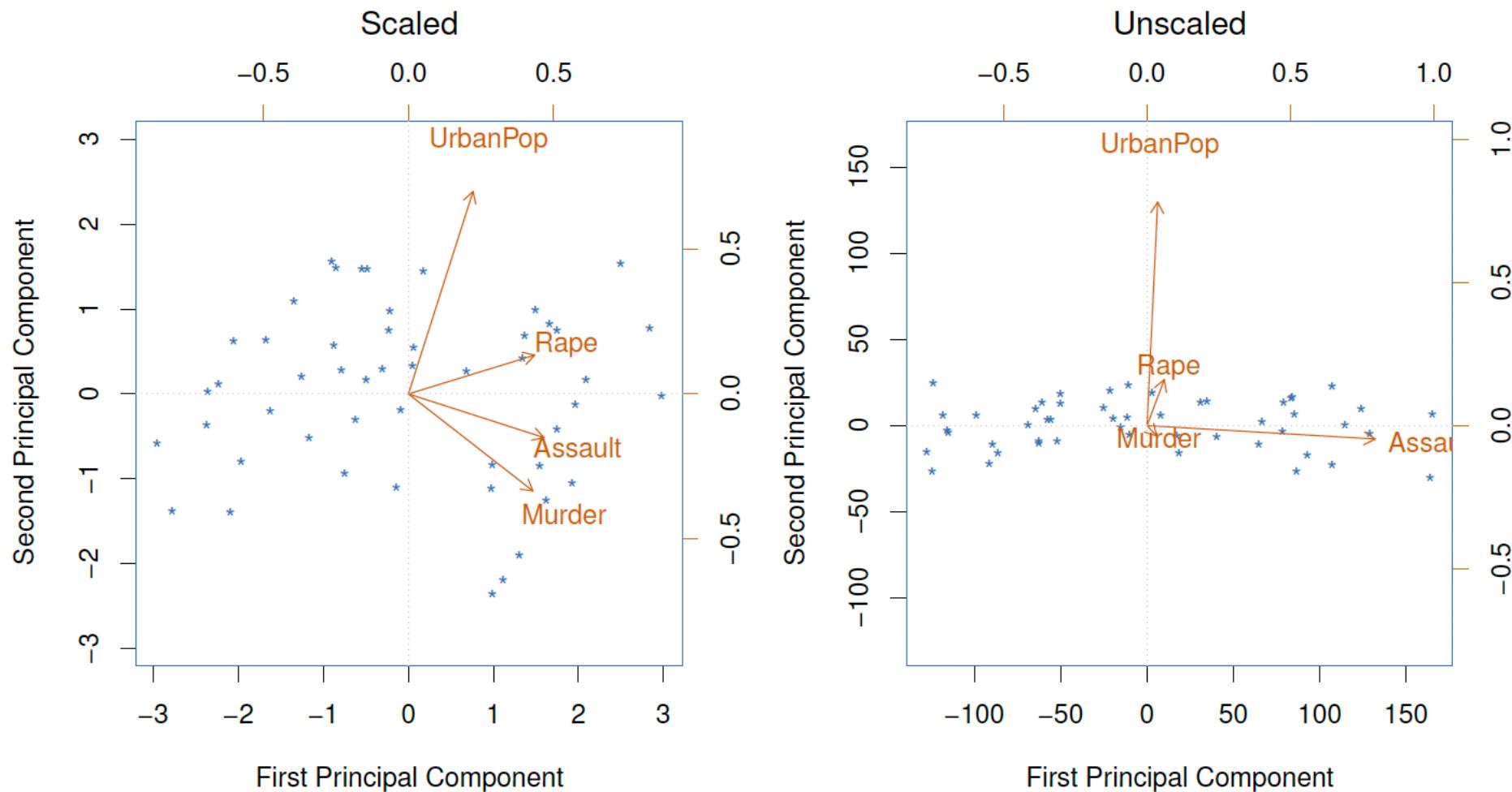
PCA: Ejemplo

*Información de
arrestos en estados
de US*

	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

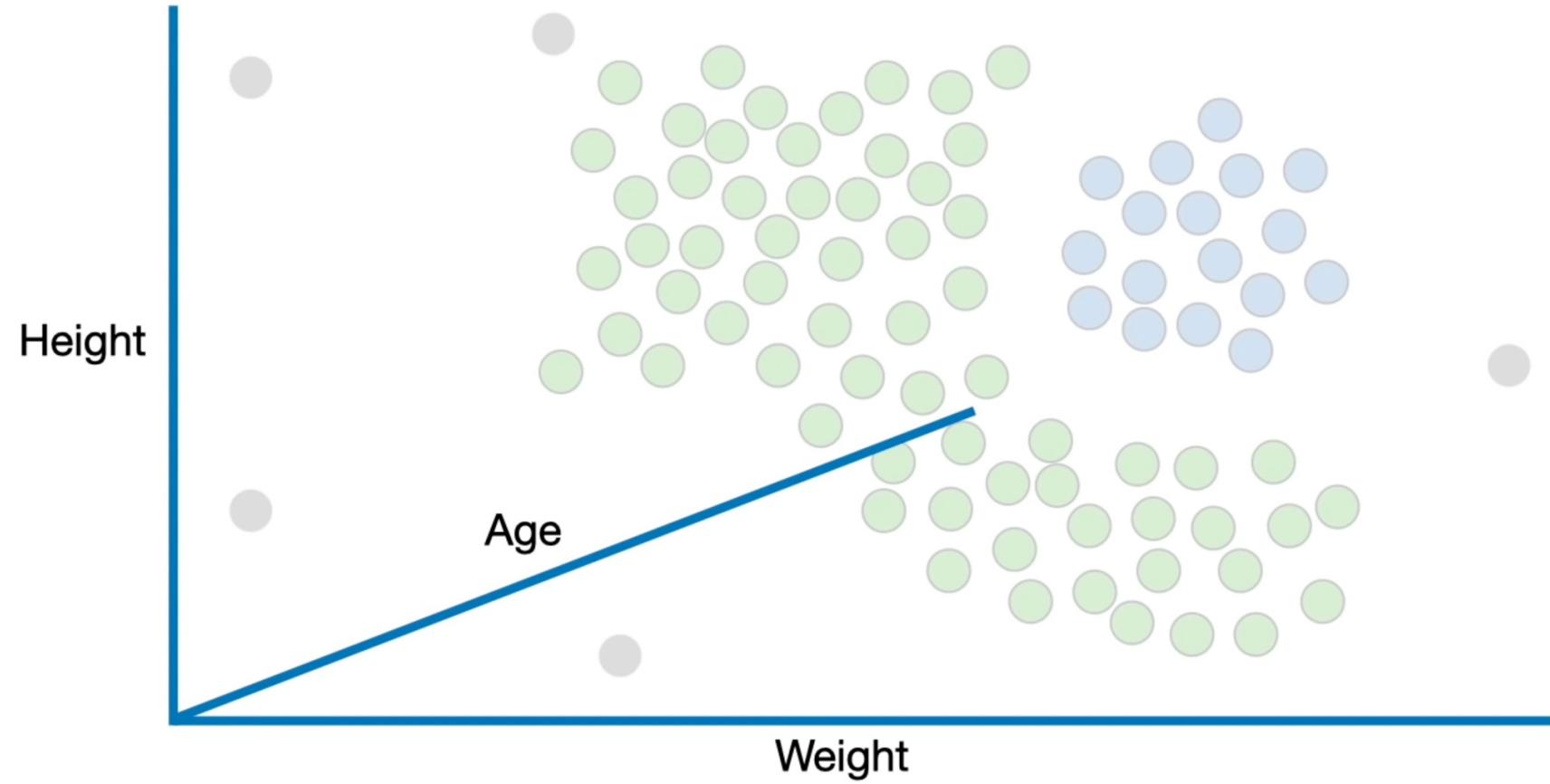


PCA: Ejemplo



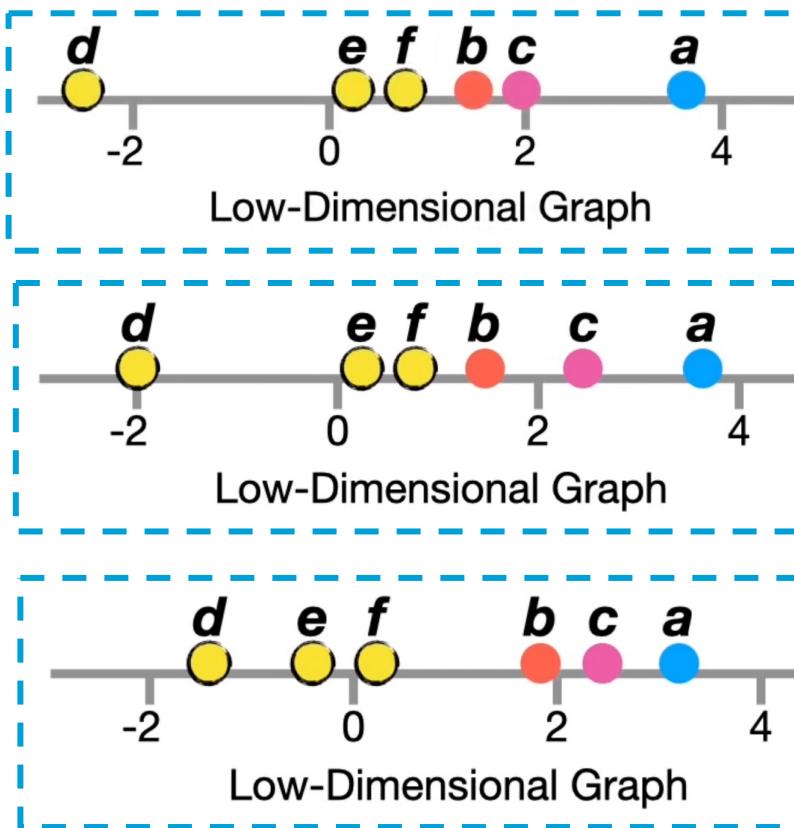
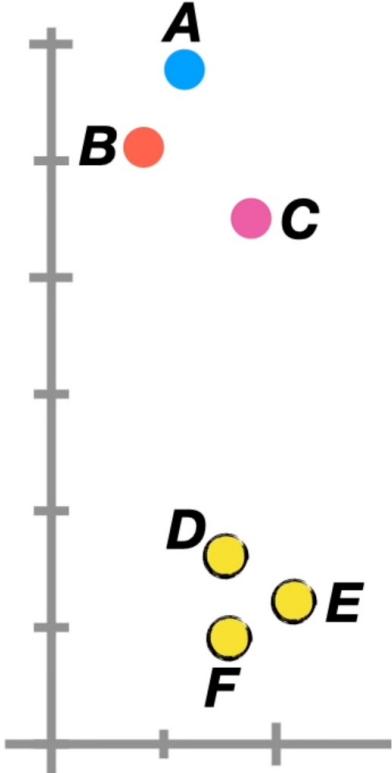
Uniform Manifold Approximation and Projection

Toma data multidimensional (3 o más dimensiones), y genera una salida como un gráfico en una baja dimensión que se puede graficar

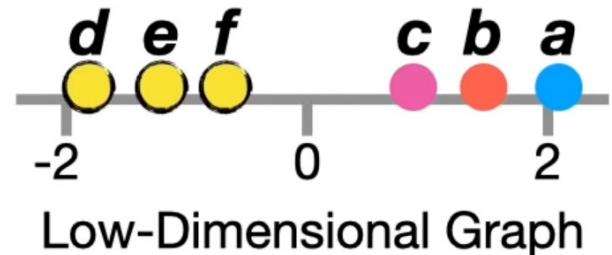


Uniform Manifold Approximation and Projection

High-Dimensional Data

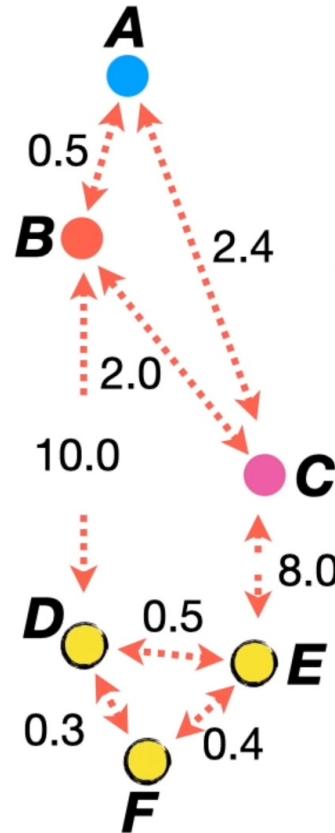


El objetivo es crear un gráfico en baja dimensión que permita preservar los grupos y las relaciones entre si



Uniform Manifold Approximation and Projection

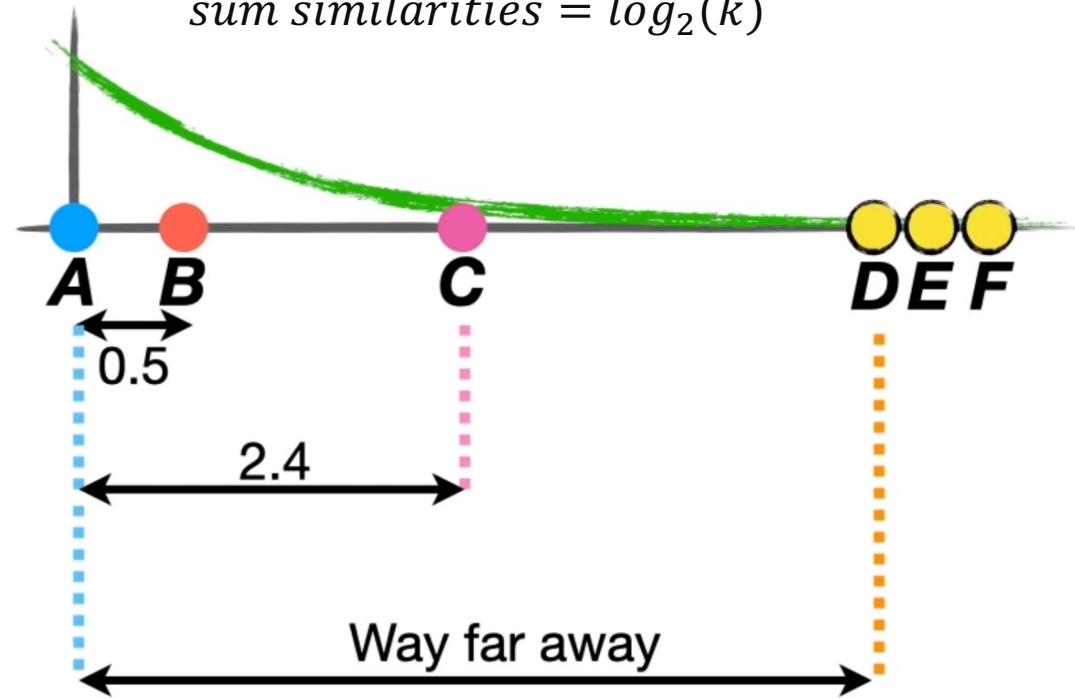
Calcula la distancia entre los pares de puntos



$$\text{Similarity Score}(A, B) = e^{-(d(A,B)-d(A,N^*))/\sigma_B}$$

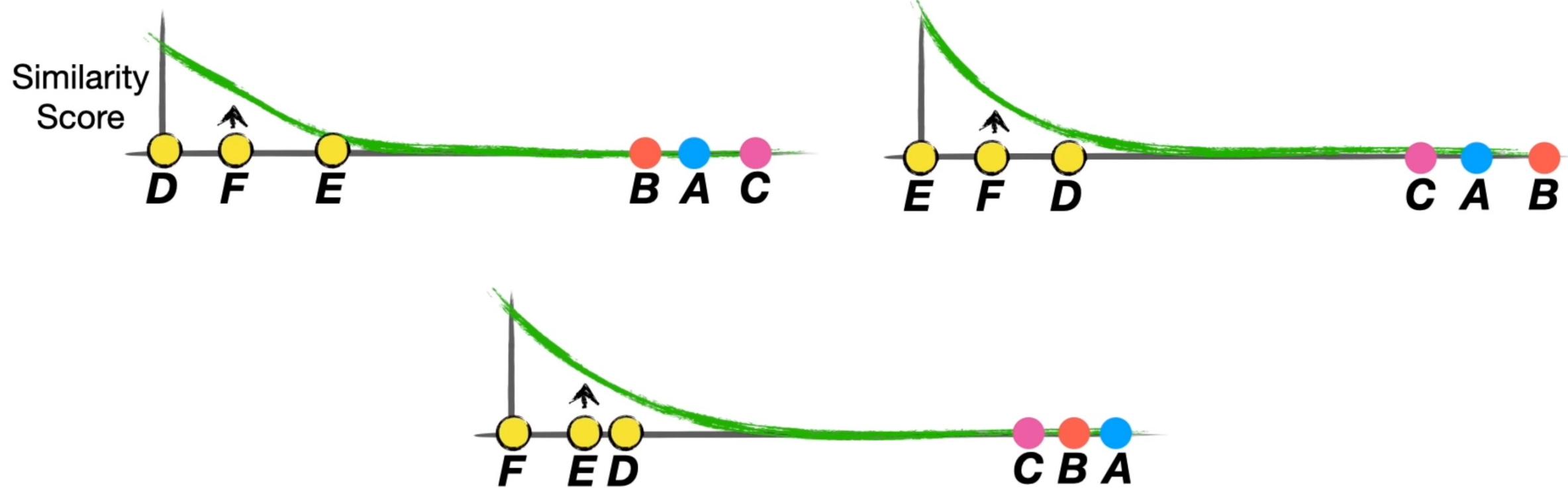
Cálculo de similaridad asociado con cada punto (y los vecinos)

$$\text{sum similarities} = \log_2(k)$$

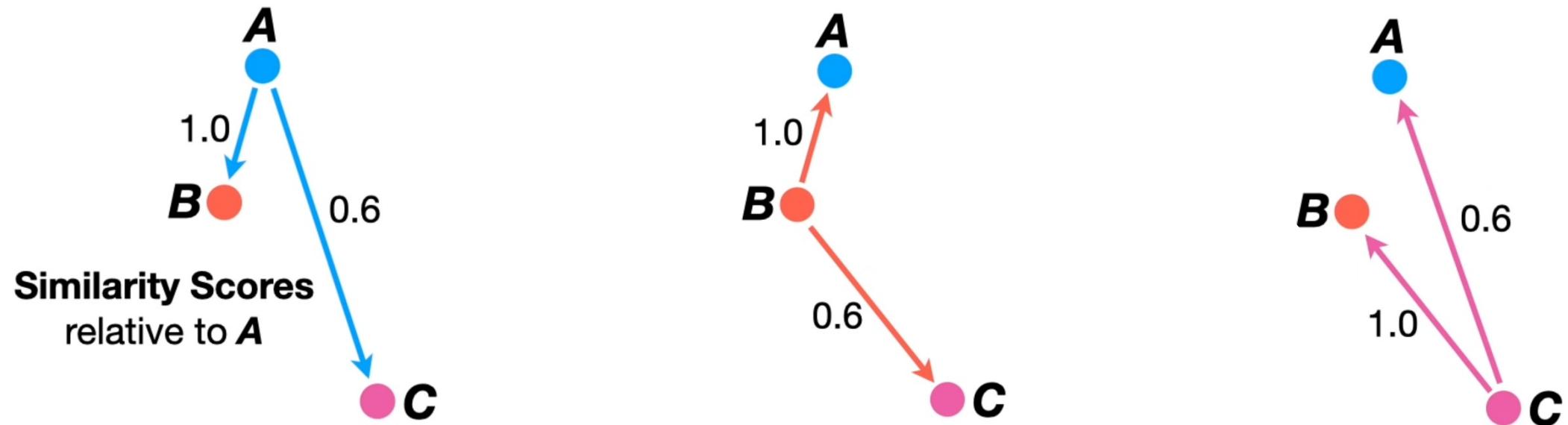


N^* corresponde al vecino más cercano

Uniform Manifold Approximation and Projection



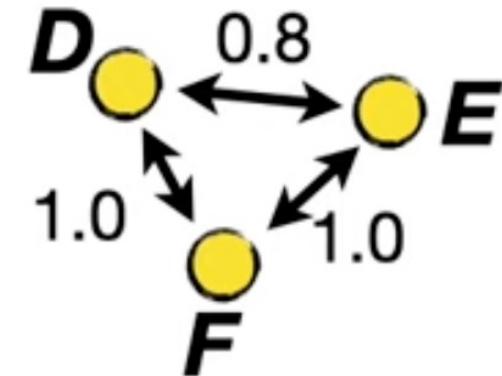
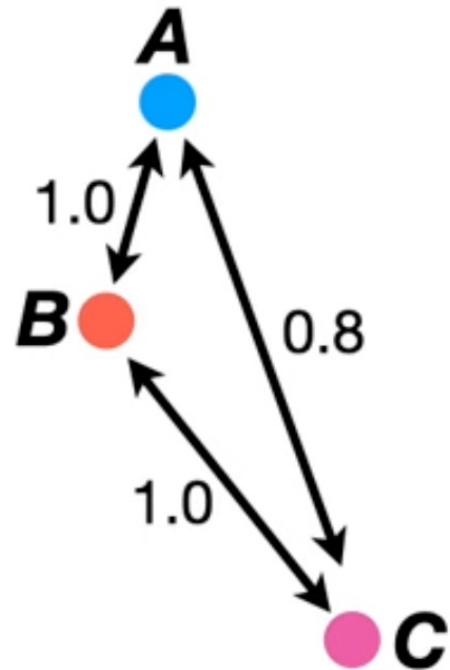
Uniform Manifold Approximation and Projection



$$\text{Symmetrical Score} = (S_1 + S_2) - S_1 S_2$$

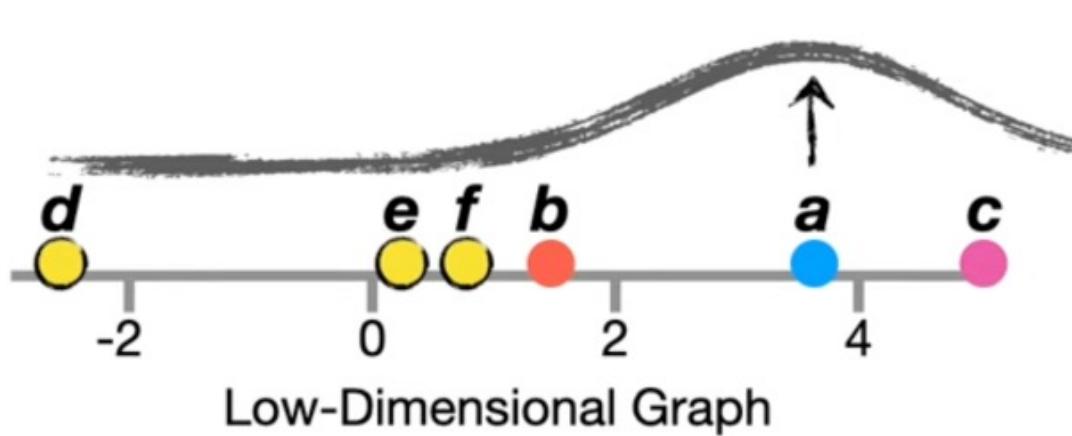
Uniform Manifold Approximation and Projection

Symmetric
Similarity Scores



Uniform Manifold Approximation and Projection

Se realiza el cálculo de los score de similaridad utilizando una distribución t



$$\text{Low-d. Scores} = \frac{1}{1 + ad^{2\beta}}$$

Alfa y beta controlan que tan juntos o separados los puntos quedan en la proyección

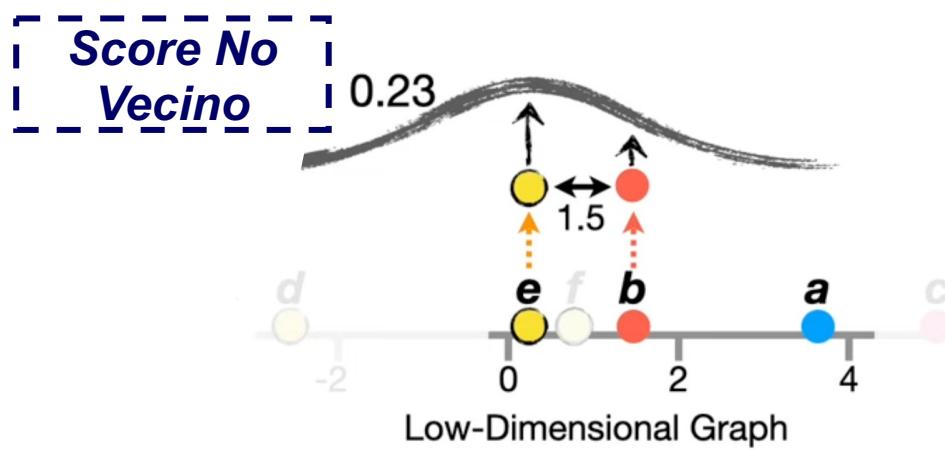
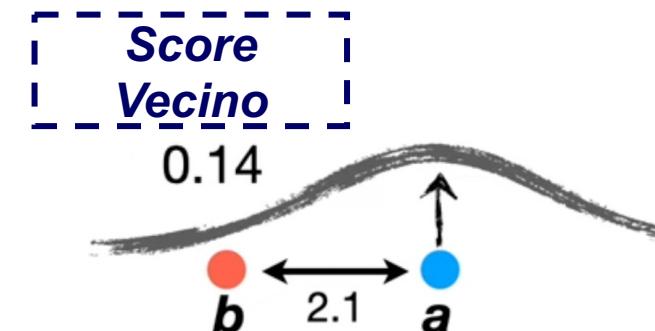
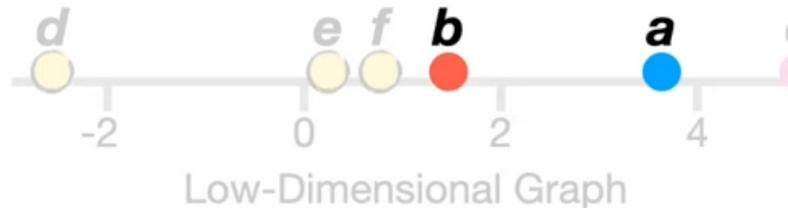
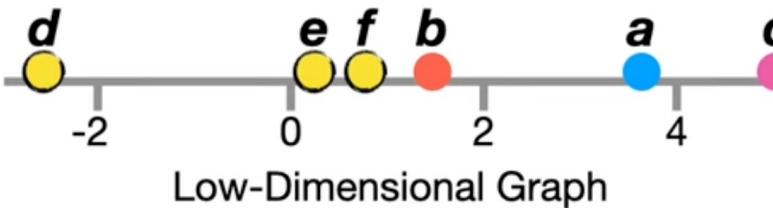
Uniform Manifold Approximation and Projection

¿Cómo reproducir los clusters de alta dimensión en baja dimensión?

1. Seleccionar dos puntos a ser acercados entre si (selección aleatoria de puntos en el mismo cluster)
2. Seleccionar aleatoriamente cual de esos puntos será movido
3. Calcular el score en baja dimensión del punto a mover respecto al punto a ser movido
4. Seleccionar un punto aleatorio de otro cluster, del cual se debe alejar y calcula el score del punto en movimiento y del punto en el cual se va a alejar
5. Dependiendo de si se mueve el punto hacia uno de los dos puntos, se mira la cercanía o lejanía de los puntos de interés y mueve el punto

Uniform Manifold Approximation and Projection

¿Cómo reproducir los clusters de alta dimensión en baja dimensión?



Movemos el punto minimizando el siguiente costo (utilizando gradiente descendiente)

$$\text{Cost} = \log\left(\frac{1}{\text{neighbor}}\right) + \log\left(\frac{1}{1 - \text{not neighbor}}\right)$$

Uniform Manifold Approximation and Projection

- ✓ Busca calcular medidas de similaridad para identificar clusters, de manera que puedan ser preservados en una dimensión menor
- ✓ Es un algoritmo relativamente rápido, incluso con conjunto de datos grandes
- ✓ Muestras similares tienden a agruparse en la salida final, lo cual es útil para identificación de grupos e incluso outliers
- ✓ Un valor pequeño de vecinos, resulta en clusters pequeños independientes, mientras que un valor grande de vecinos, puede modelar dependencias más generales pero no específicas en los datos (parámetro a ajustar)

Estrategias de Segmentación

En estrategias de mercadeo, se tienen muchos clientes, pero solo hay algunas categorías principales en las que se pueden agrupar la mayoría de tus clientes.

- ✓ Cazador de ofertas
- ✓ Hombre o mujer con una misión
- ✓ Comprador impulsivo
- ✓ Padre cansado
- ✓ DINK (Doble ingreso, sin hijos)



Estrategias de Segmentación

En estrategias de ubicación de tiendas, se quieren abrir nuevas tiendas de comestibles en los Estados Unidos basándote en la demografía. ¿Dónde deberías ubicar los siguientes tipos de nuevas tiendas?

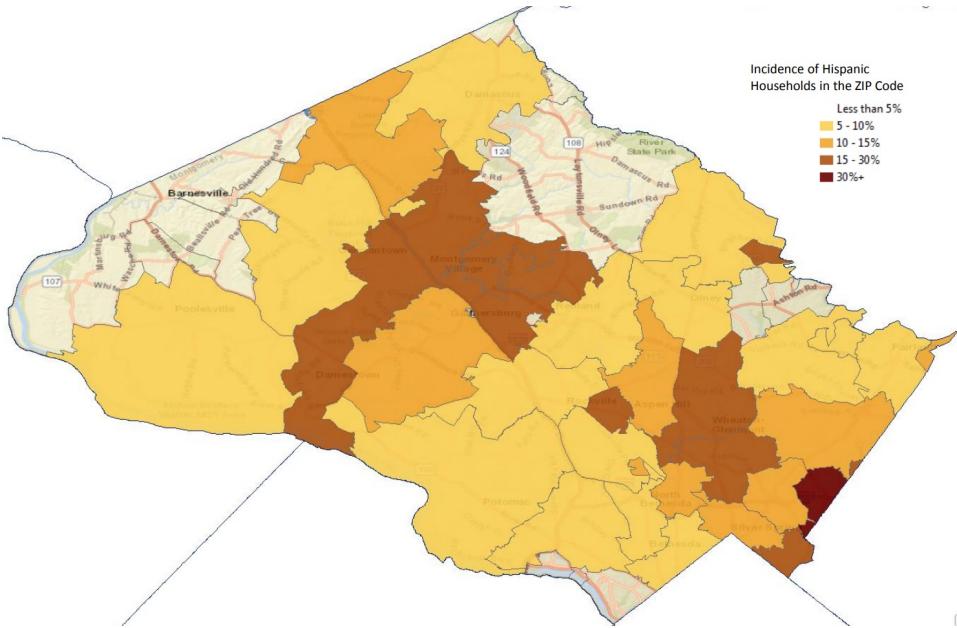
- ✓ Tiendas de comestibles económicas de bajo costo
- ✓ Pequeñas tiendas de comestibles boutique
- ✓ Grandes supermercados de servicio completo



Estrategias de Segmentación

EurekaFacts **Segmentos**

Hispanic communities understood

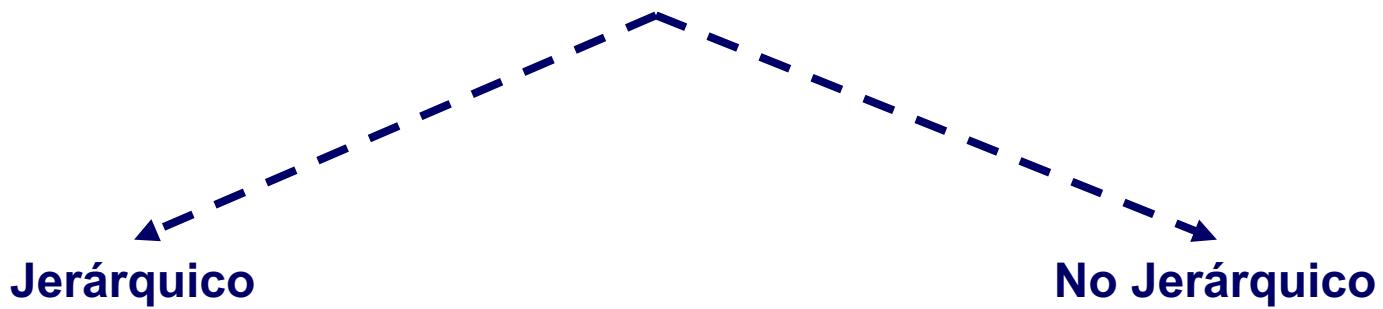


Examples of Segmentos Data Usages

1. **Choose a zip code, city, metro area or region** and Segmentos will provide a very complete picture of the Latino population that lives there and the various distinct groups that make up the Latino community including how the area or region differs from the national population.
2. **Choose a demographic profile** and Segmentos will inform you of the complete consumer profile including the size of your target demographic and the communities where you can find them.
3. **Choose a particular attribute that relates to your objectives** (e.g. large family size, renters, millennials, etc.) and Segmentos will provide you with Latino groups and where they are concentrated across the US.
4. **Match a list of customers** to the Segmentos clusters, enriching the information you have on your customers and offering insights on how to find more customers that look like those you have already won over.
5. **Add the Segmentos database** to your custom segmentation and customer profiling system to serve as a stream of attributes informing your community profiles.
6. **Use geo-location** to better understand your customer/client's community and implement geo-intelligent strategies that resonate with each market.
7. **Contact us to develop your own customized segmentation approach.** Layering your own data and custom new datasets ,we can develop highly effective market and customer profiles and manage a data-driven platform to engage with your Latino market.

Clustering

El análisis cluster se ocupa de buscar patrones en una población para agrupar observaciones (multivariadas) en conjuntos homogéneos, los cuales se denominan clusters. Al interior de cada cluster las observaciones u objetos son semejantes, sin embargo, los clusters son diferentes entre sí (Rencher, 2002)



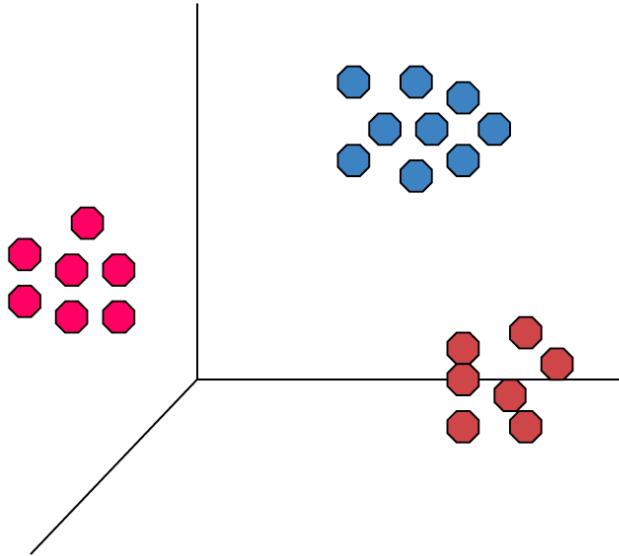
No se conoce cuantos grupos se desean, pero se termina con una representación visual de las observaciones agrupadas

Buscamos una partición de las observaciones en un número predefinido de grupos

Objetivo de la agrupación

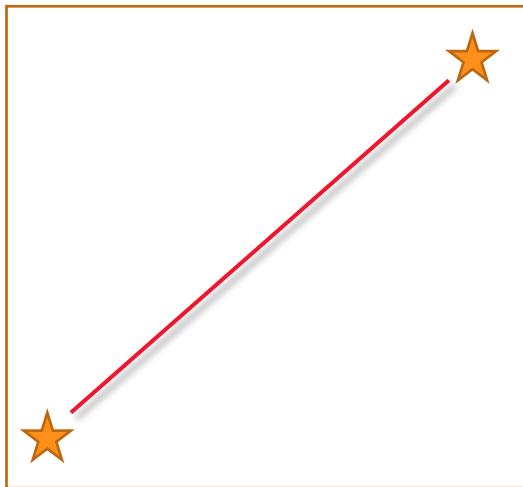
Distancias Intracluster
son minimizadas

Distancias Intercluster
son maximizadas

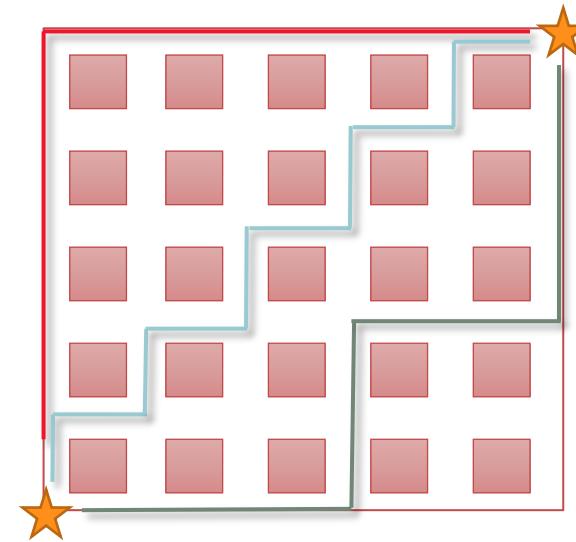


¿Cómo medimos?

“¿Cuál es la distancia más corta entre dos puntos?”



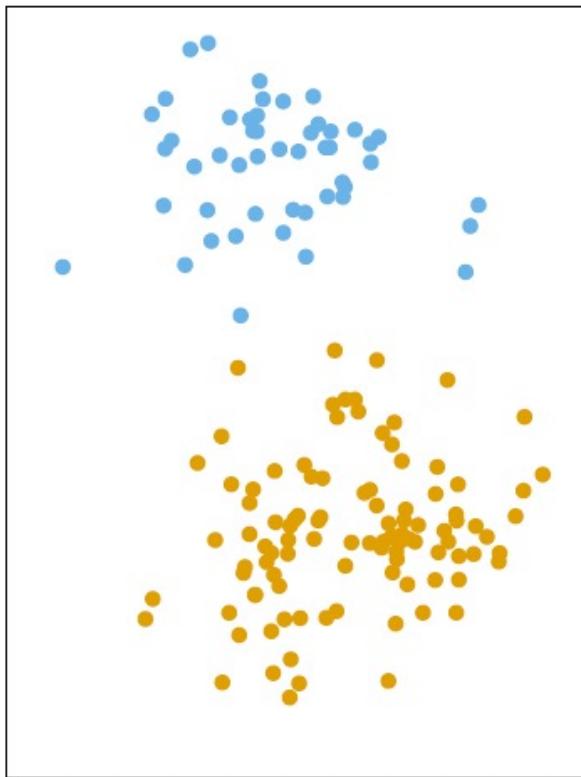
$$d_2(p, q) = \|p - q\|_2 = \sqrt{\sum_i (p_i - q_i)^2}$$



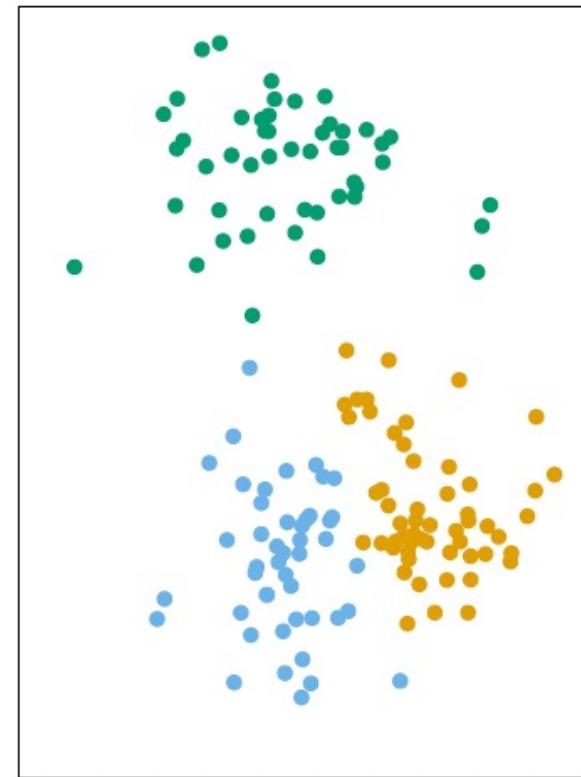
$$d_1(p, q) = \|p - q\|_1 = \sum_i |p_i - q_i|$$

¿Cuántos grupos debo considerar?

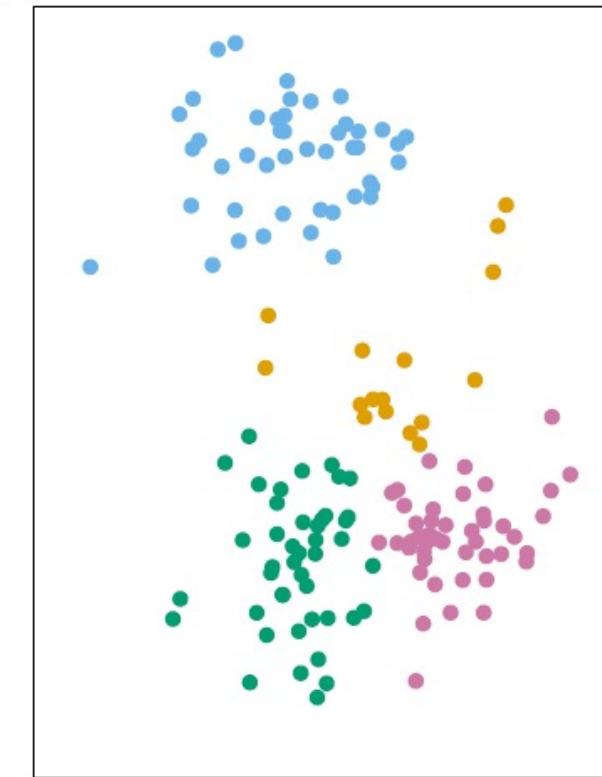
K=2



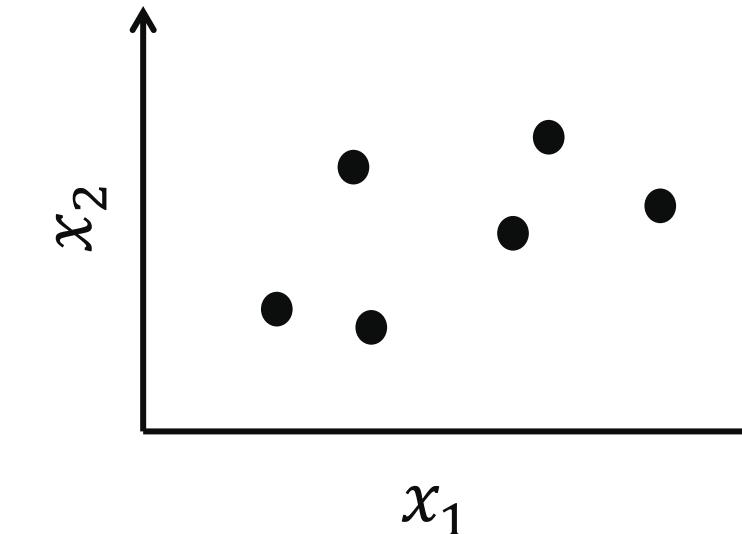
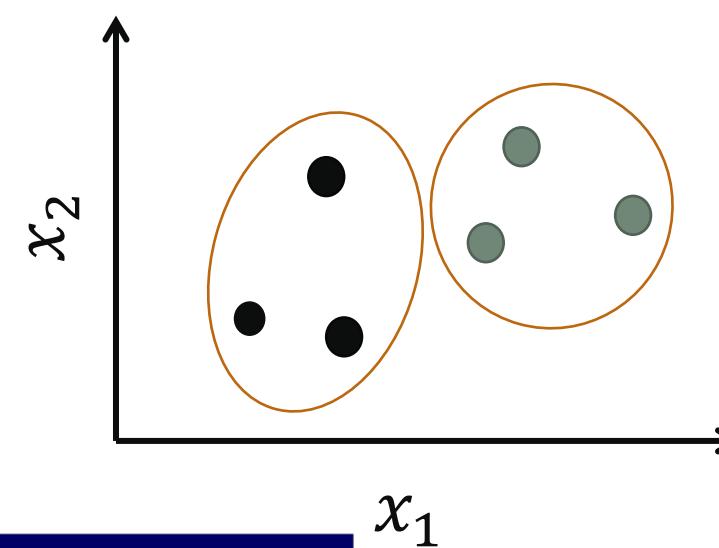
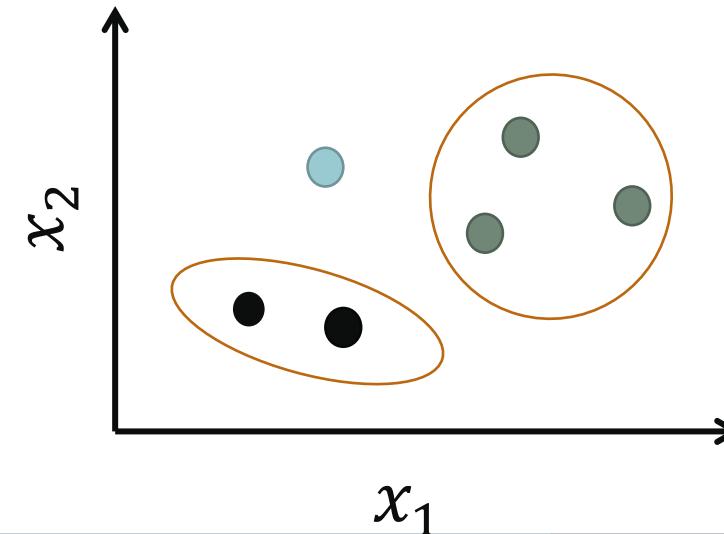
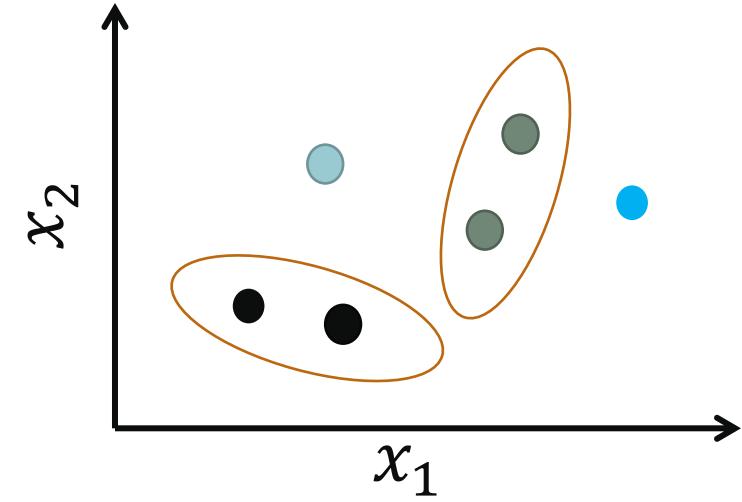
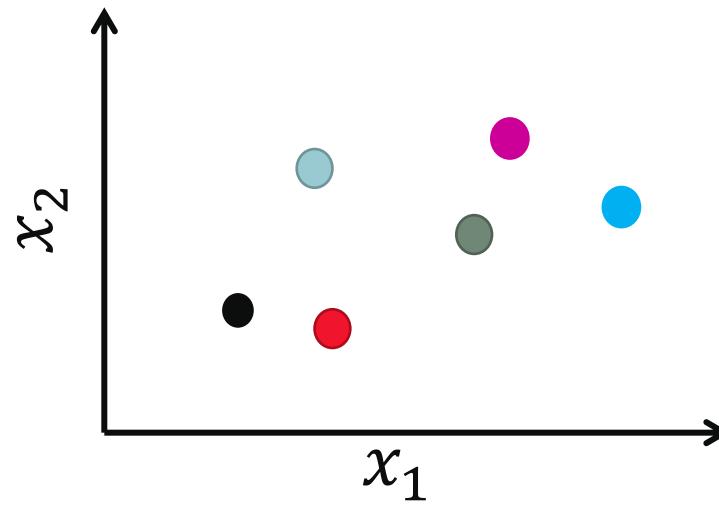
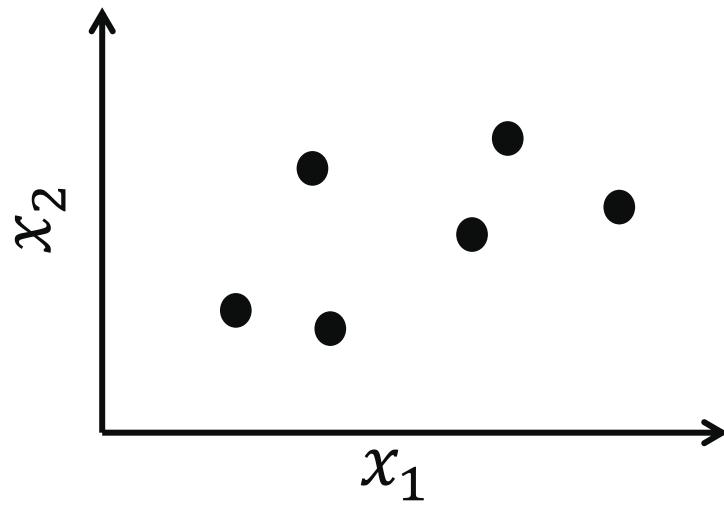
K=3



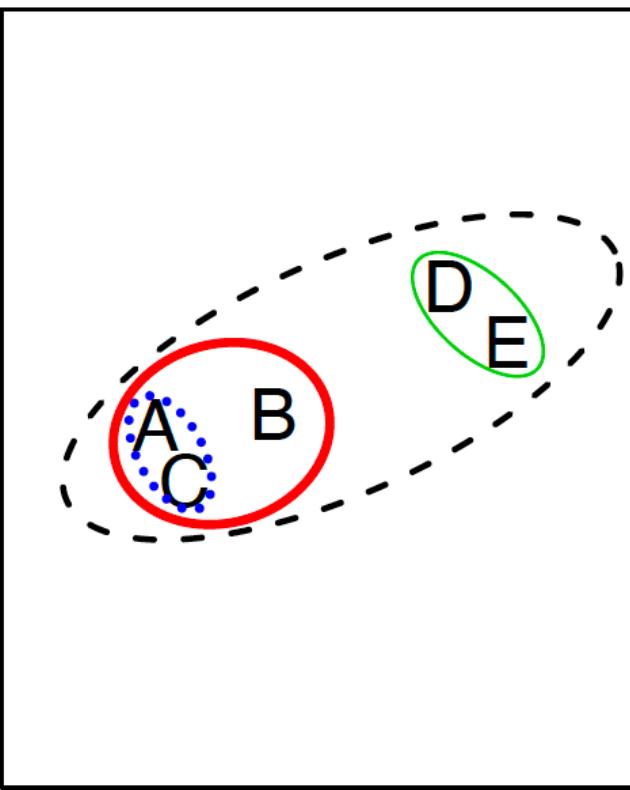
K=4



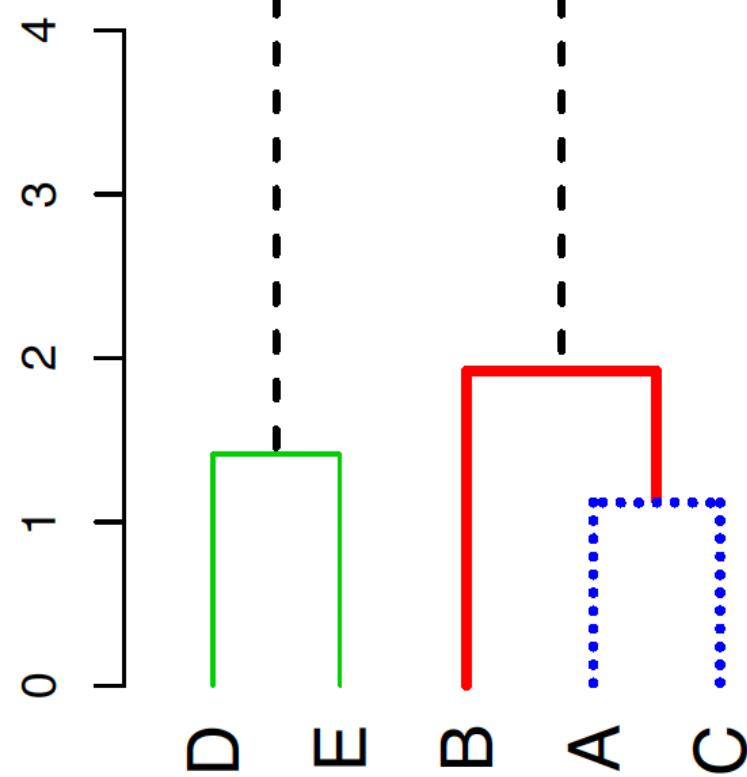
Cluster Jerárquico



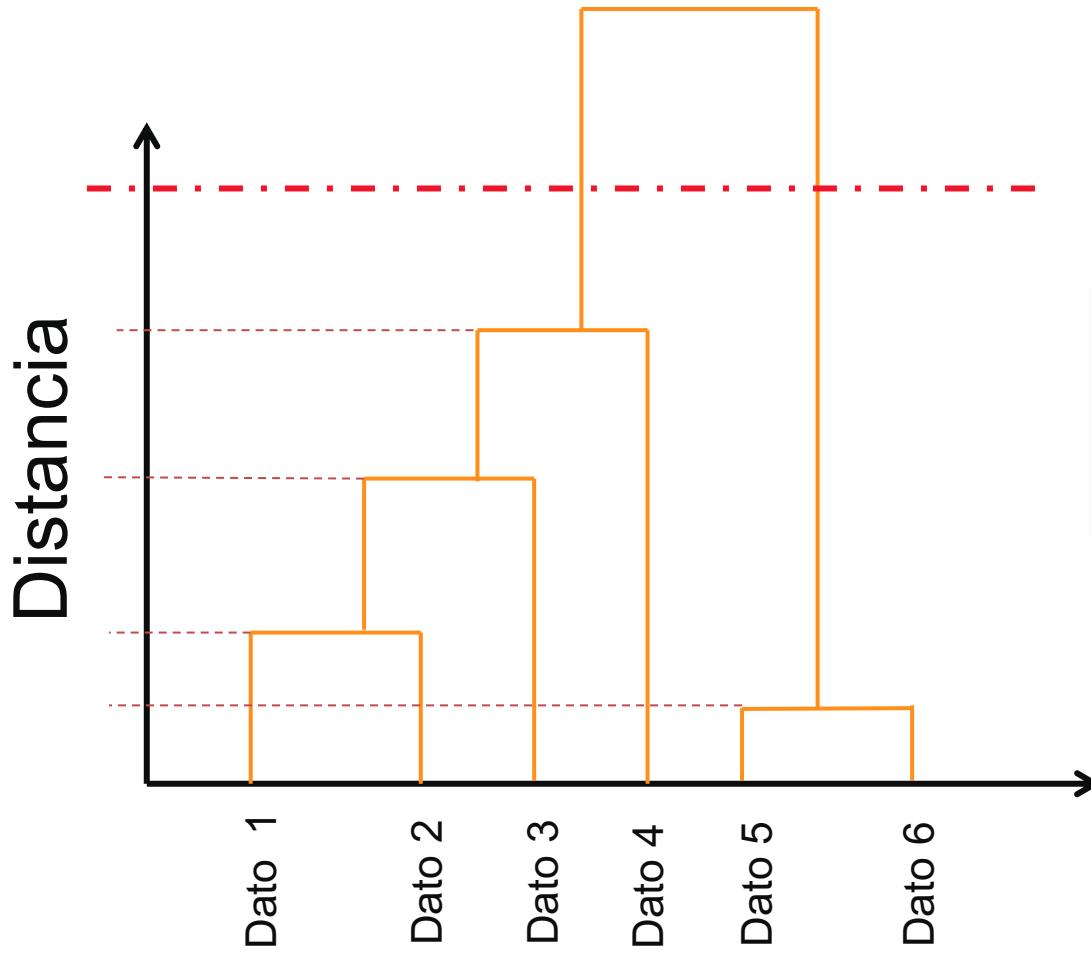
Cluster Jerárquico



Dendrogram



Dendograma



- ¿Cuándo parar?
- ¿Qué grupos unir/agregar?

Ciudad	Asesinato	Violacion	Robo	Asalto	Robo Casas	Hurto	Autos
Atlanta	16.5	24.8	106	147	1112	905	494
Boston	4.2	13.3	122	90	982	669	954
Chicago	11.6	24.7	340	242	808	609	645
Dallas	18.1	34.2	184	293	1668	901	602
Denver	6.9	41.5	173	191	1534	1368	780
Detroit	13	35.7	477	220	1566	1183	788

	Atlanta	Boston	Chicago	Dallas	Denver
Atlanta	0				
Boston	536.6419	0			
Chicago	516.37	447.4033	0		
Dallas	590.1753	833.0708	924.0035	0	
Denver	693.5741	914.9784	1073.395	527.6673	0
Detroit	716.1962	881.0858	971.5271	464.4677	358.6654

Grupo 1: Detroit+Denver

Ciudad	Asesinato	Violacion	Robo	Asalto	Robo Casas	Hurto	Autos
Atlanta	16.5	24.8	106	147	1112	905	494
Boston	4.2	13.3	122	90	982	669	954
Chicago	11.6	24.7	340	242	808	609	645
Dallas	18.1	34.2	184	293	1668	901	602
Denver	6.9	41.5	173	191	1534	1368	780
Detroit	13	35.7	477	220	1566	1183	788

	Atlanta	Boston	Chicago	Dallas
Atlanta	0			
Boston	536.6419	0		
Chicago	516.37	447.4033	0	
Dallas	590.1753	833.0708	924.0035	0
Grupo 1	693.5741	881.0858	971.5271	464.4677

Grupo 1: Detroit+Denver

Grupo 2: Chicago+Boston

Ciudad	Asesinato	Violacion	Robo	Asalto	Robo Casas	Hurto	Autos
Atlanta	16.5	24.8	106	147	1112	905	494
Boston	4.2	13.3	122	90	982	669	954
Chicago	11.6	24.7	340	242	808	609	645
Dallas	18.1	34.2	184	293	1668	901	602
Denver	6.9	41.5	173	191	1534	1368	780
Detroit	13	35.7	477	220	1566	1183	788

	Atlanta	Dallas
Atlanta	0	
Grupo 2	516.37	
Dallas	590.1753	0
Grupo 1	693.5741	464.4677

Grupo 1: Detroit+Denver

Grupo 2: Chicago+Boston

Grupo 3: Grupo 1+Dallas

Ciudad	Asesinato	Violacion	Robo	Asalto	Robo Casas	Hurto	Autos
Atlanta	16.5	24.8	106	147	1112	905	494
Boston	4.2	13.3	122	90	982	669	954
Chicago	11.6	24.7	340	242	808	609	645
Dallas	18.1	34.2	184	293	1668	901	602
Denver	6.9	41.5	173	191	1534	1368	780
Detroit	13	35.7	477	220	1566	1183	788



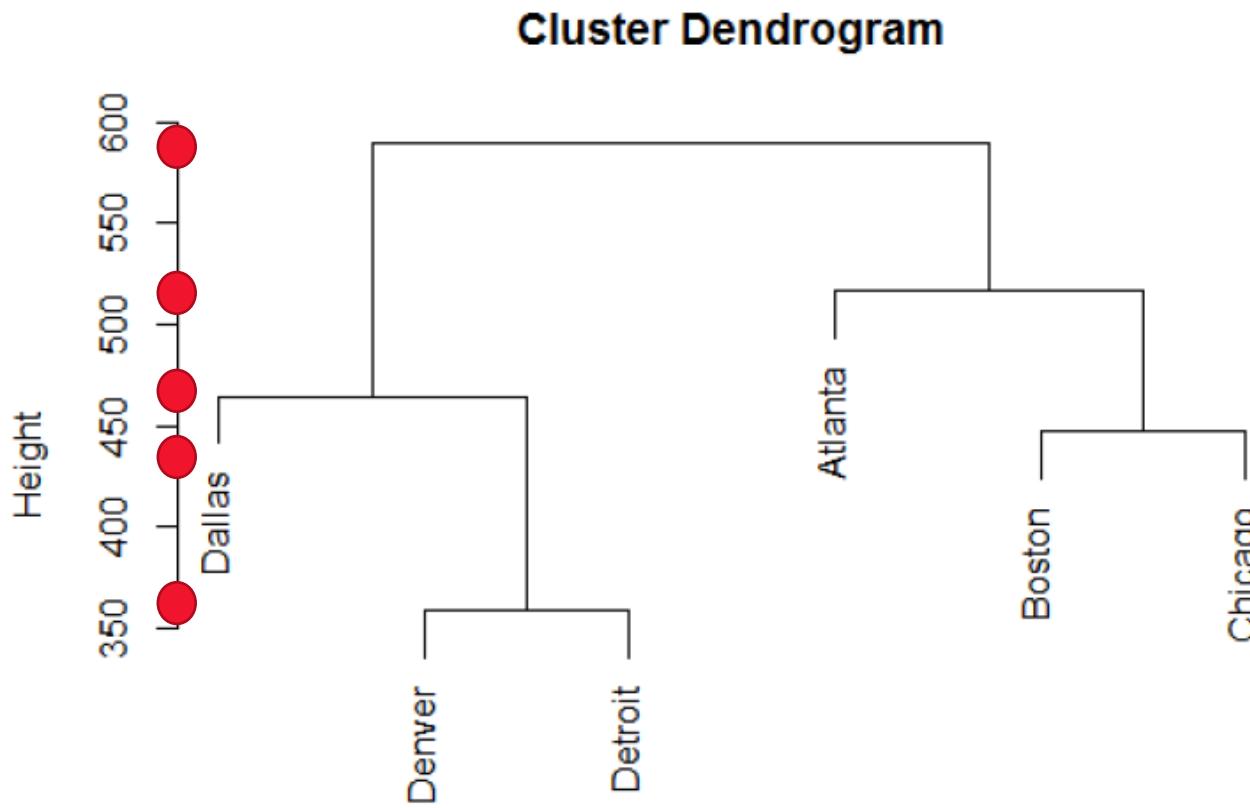
Grupo 1: Detroit+Denver

Grupo 2: Chicago+Boston

Grupo 3: Grupo 1+Dallas

Grupo 4: Grupo 2+Atlanta

- Grupo 1:** Detroit+Denver (358.6)
Grupo 2: Chicago+Boston (447.4)
Grupo 3: Grupo 1+Dallas (464.4)
Grupo 4: Grupo 2+Atlanta (516)



Dendograma

Posibles modificaciones (Considere A y B clusters):

✓ **Vecino más cercano**

$$D(A, B) = \min\{d(y_i, y_j) | y_i \in A, y_j \in B\}$$

✓ **Vecino más lejano**

$$D(A, B) = \max\{d(y_i, y_j) | y_i \in A, y_j \in B\}$$

✓ **Agrupamiento promedio**

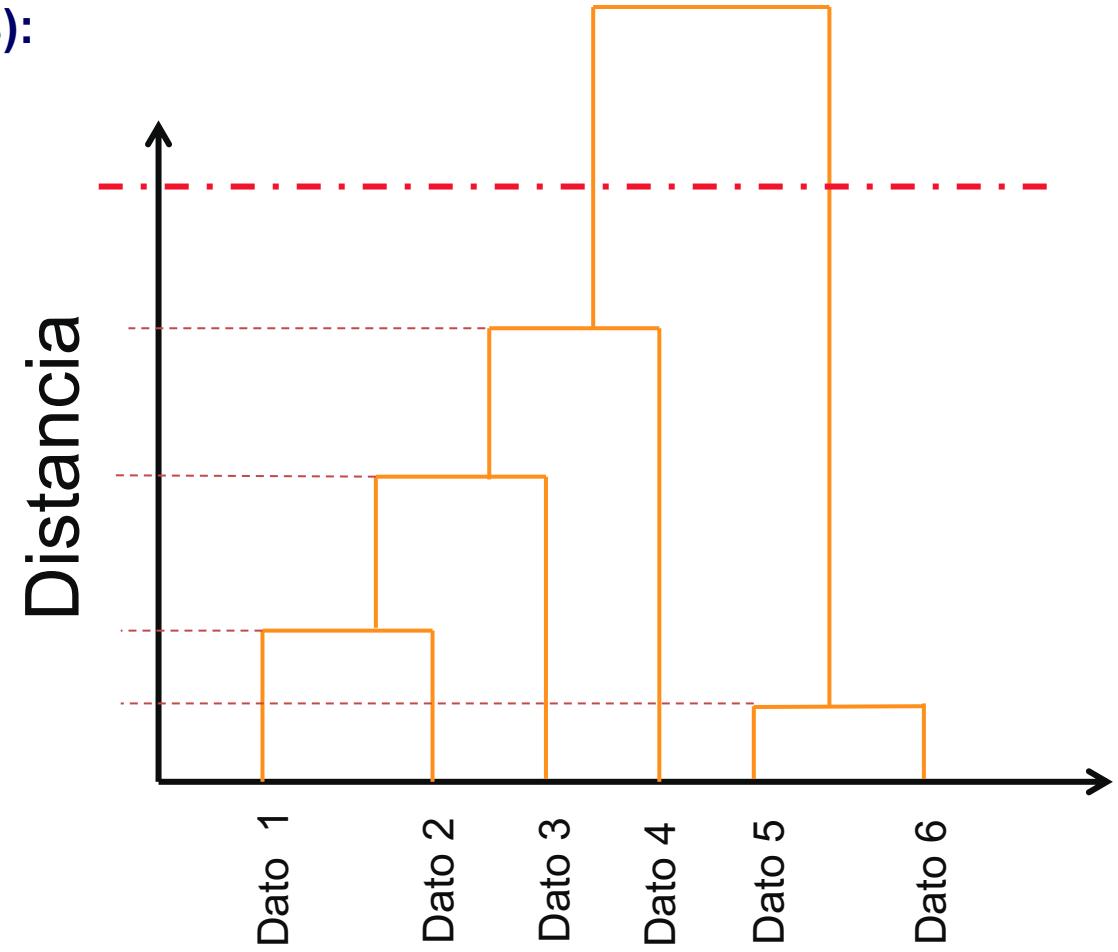
$$D(A, B) = \frac{1}{n_a n_b} \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} d(y_i, y_j)$$

✓ **Método del centroide**

$$D(A, B) = d(\bar{Y}_A, \bar{Y}_B) \quad \bar{Y}_{AB} = \frac{n_A \bar{Y}_A + n_B \bar{Y}_B}{n_A + n_B}$$

✓ **Método de la mediana**

$$m_{AB} = \frac{\bar{Y}_A + \bar{Y}_B}{2}$$



K-Means

Sea C_1, C_2, \dots, C_k los conjuntos donde se agrupan las observaciones en cada cluster. Estos conjuntos satisfacen las siguientes propiedades:

- ✓ $C_1 \cup C_2 \cup \dots \cup C_k = \{1, \dots, n\}$. Cada observación pertenece a al menos uno de los K clusters posibles
- ✓ $C_i \cap C_j = \emptyset \forall i \neq j$. Los clusters no se solapan, es decir, que una observación no puede pertenecer a más de un cluster

K-Means

- ✓ La idea dentro del método de K-Means es que una buena agrupación es aquella en la que la variación en cada cluster es tan pequeña como sea posible
- ✓ La variación en cada cluster es una medida de que tanto las observaciones de un cluster difieren de las demás
- ✓ Se busca resolver el problema:

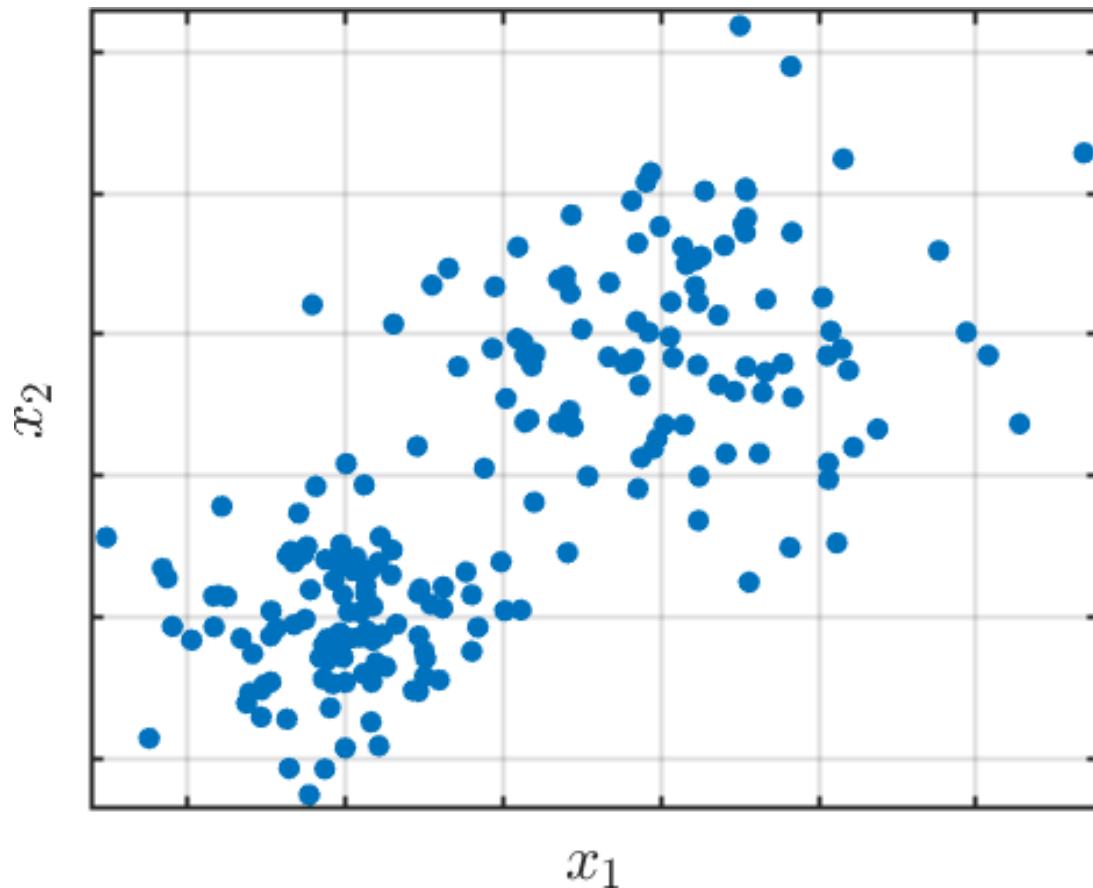
$$\underset{C_1, \dots, C_K}{\text{minimize}} \left\{ \sum_{k=1}^K \text{WCV}(C_k) \right\}$$

K-Means

$$\text{WCV}(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

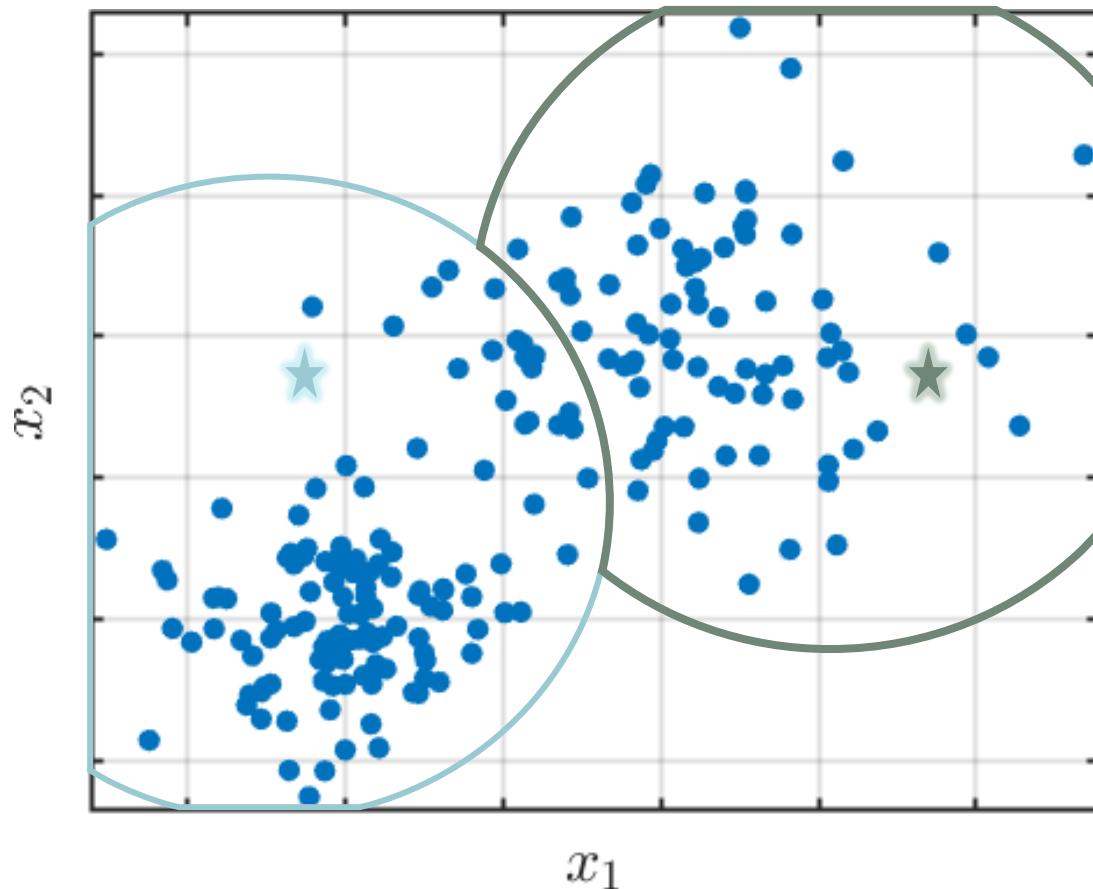
$$\underset{C_1, \dots, C_K}{\text{minimize}} \left\{ \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \right\}$$

K-Means



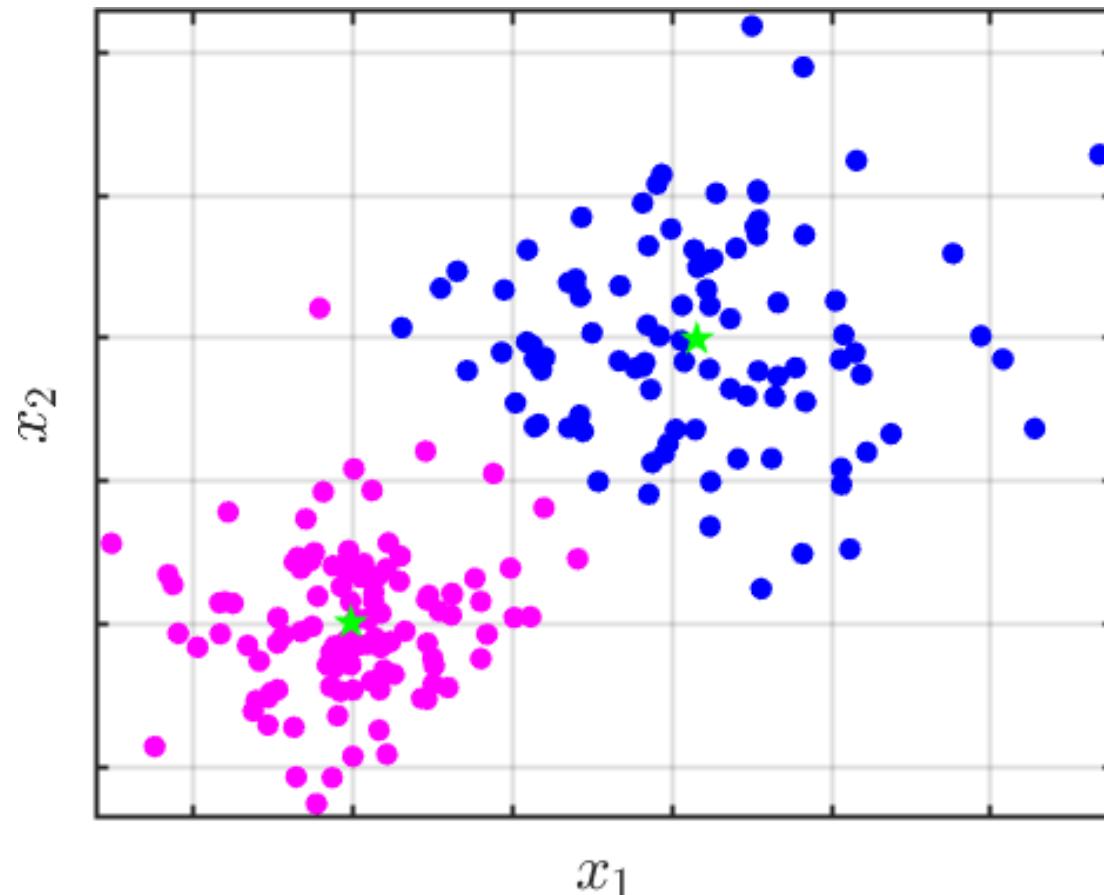
- ¿Cuántos grupos esperamos? (K)

K-Means



- ¿Cuántos grupos esperamos? (K)
- Iniciar los centros de los grupos aleatoriamente.
- Calcular la distancia de cada dato a los centros.

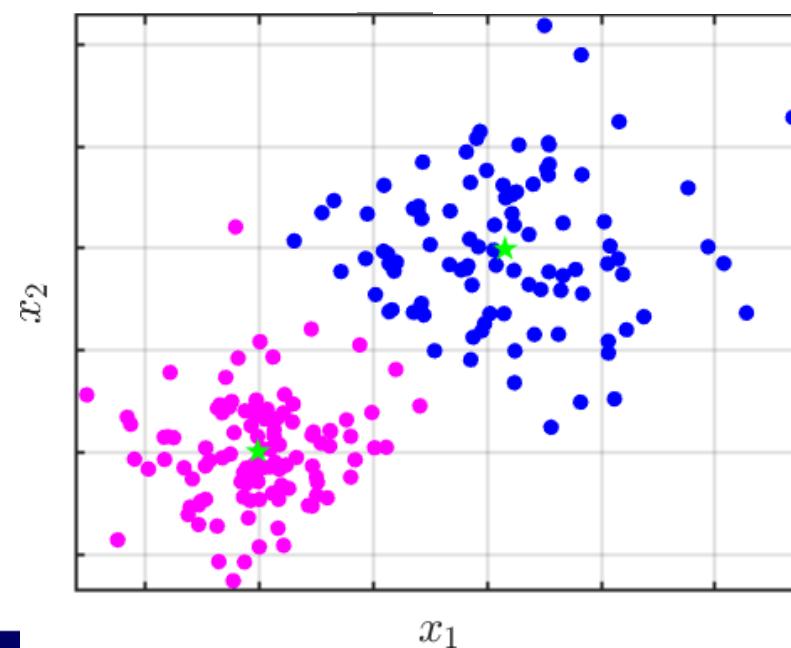
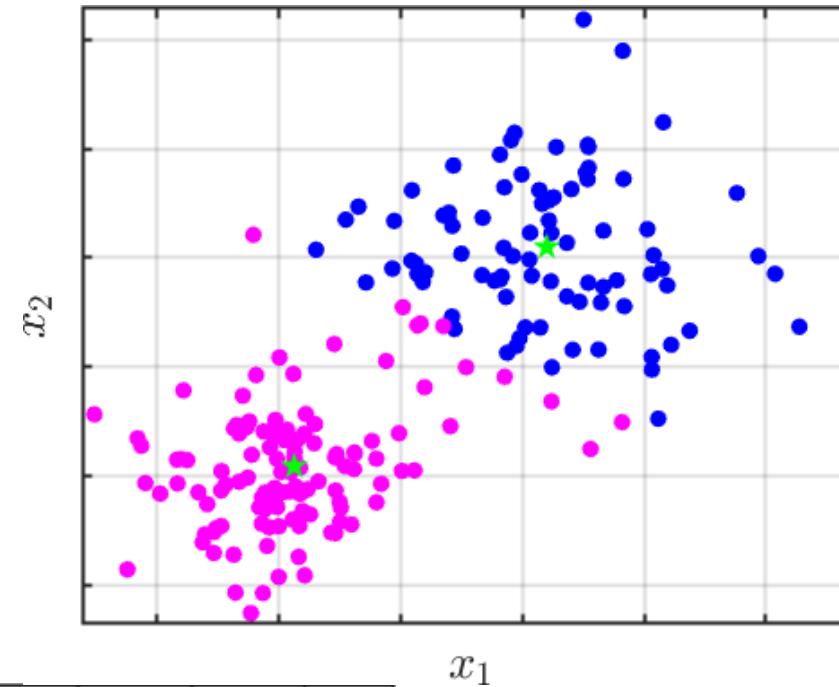
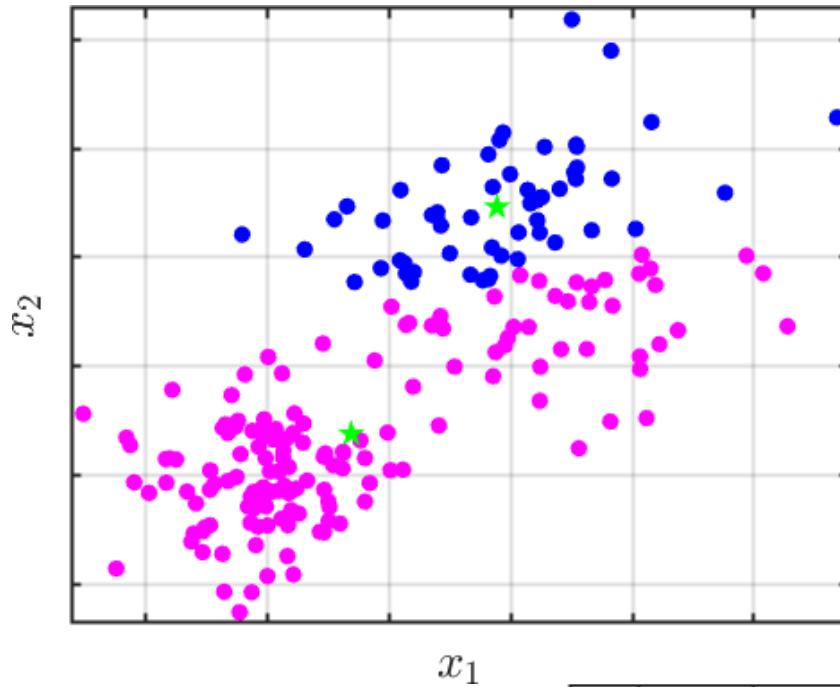
K-Means



- ¿Cuántos grupos esperamos? (K)
- Iniciar los centros de los grupos aleatoriamente.
- Calcular la distancia de cada dato a los centros.
- El centroide de cada grupo es el nuevo centro.

Nuevo centro:

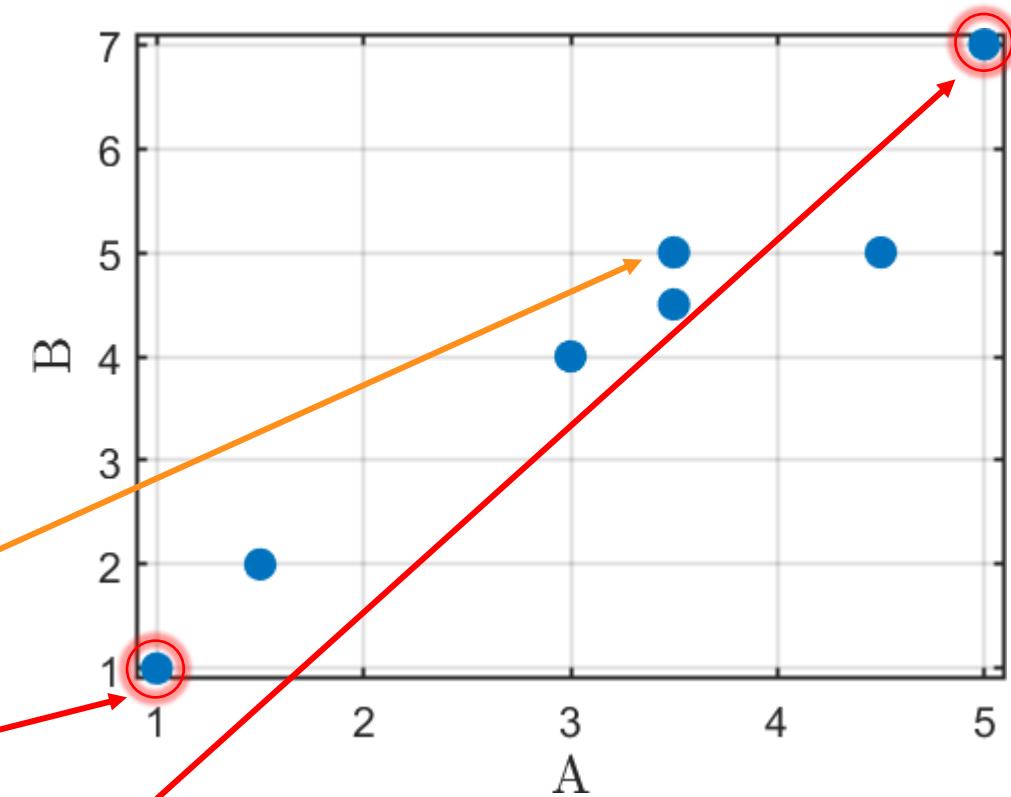
$$c_k = \frac{1}{|G_k|} \sum x_i$$



K-Means: Ejemplo

Queremos agrupar el siguiente grupo de datos en **2** grupos.

Dato	Característica A	Característica B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5



Inicialmente: Centro1: (1 , 1) y Centro2 (5, 7)

Dato	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Dist C1	Dist C2
0	7.2
1.1	6.1
??	??
7.2	0
4.7	2.5
??	??
4.3	2.9

Grupo
1
1
1
2
2
2
2

Dato	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0

Dato	A	B
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

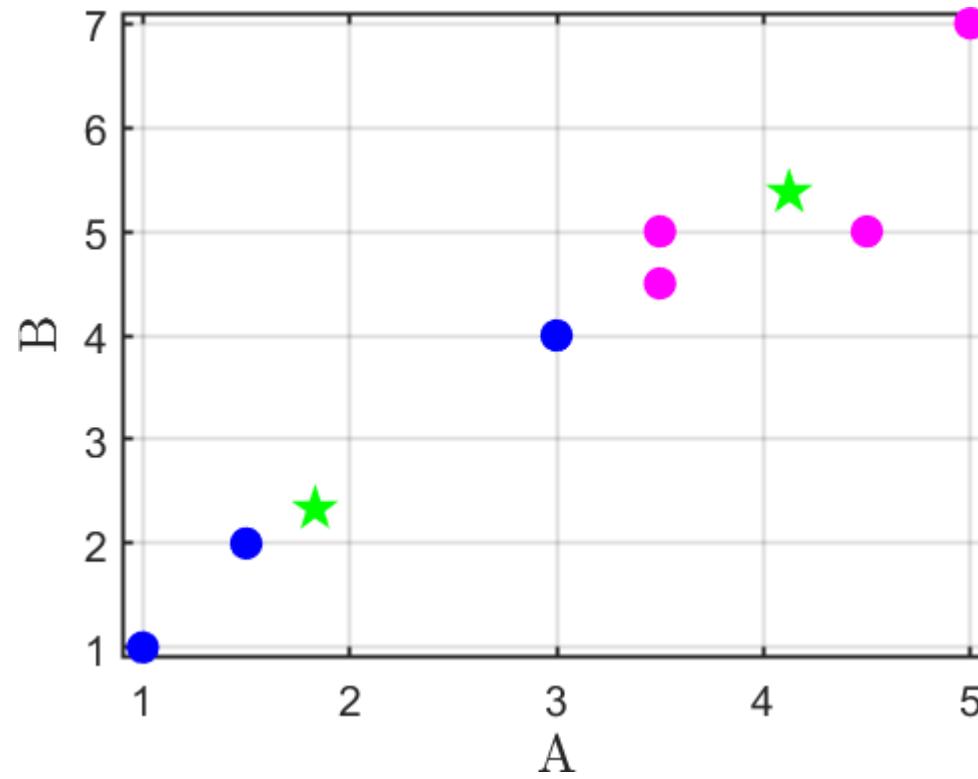
Nuevo centro:

$$C1 = \frac{1}{|G1|} \sum x_i$$

Dato	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0

Dato	A	B
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

	A	B
C1	1.8	2.3
C2	4.1	5.3



Dato	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Dist C1	Dist C2
1.6	5.4
0.5	4.3
??	??
5.6	1.8
3.1	0.7
??	??
2.7	1.1

Grupo
1
1
2
2
2
2
2

Dato	A	B
1	1.0	1.0
2	1.5	2.0

Dato	A	B
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

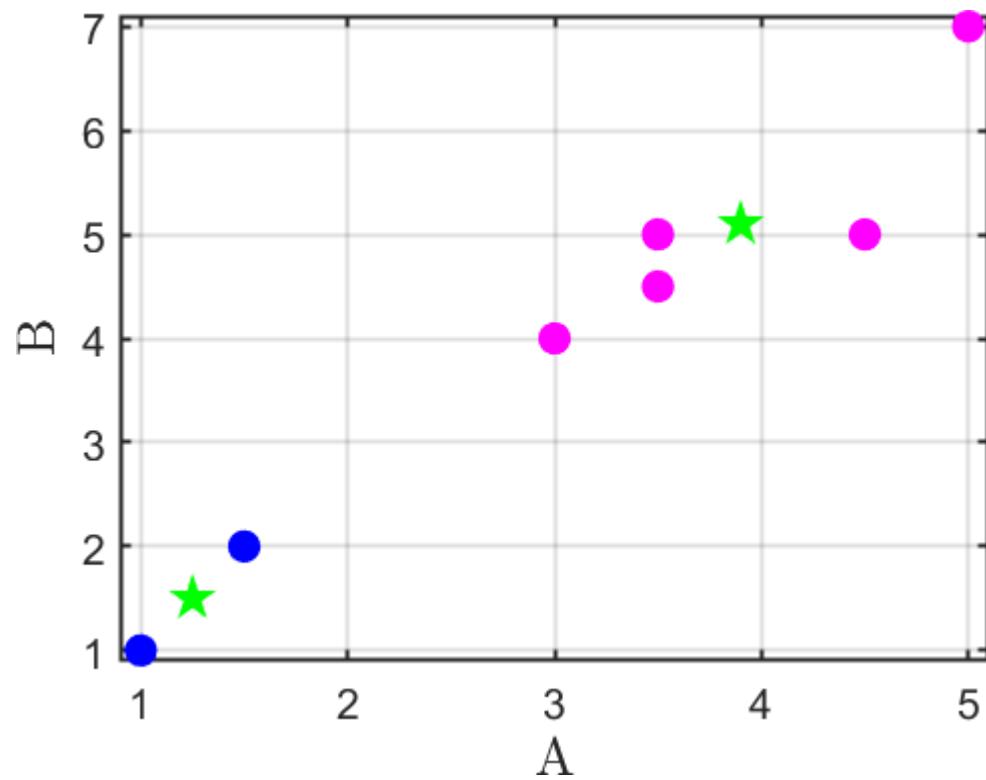
Nuevo centro:

$$C1 = \frac{1}{|G1|} \sum x_i$$

Dato	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0

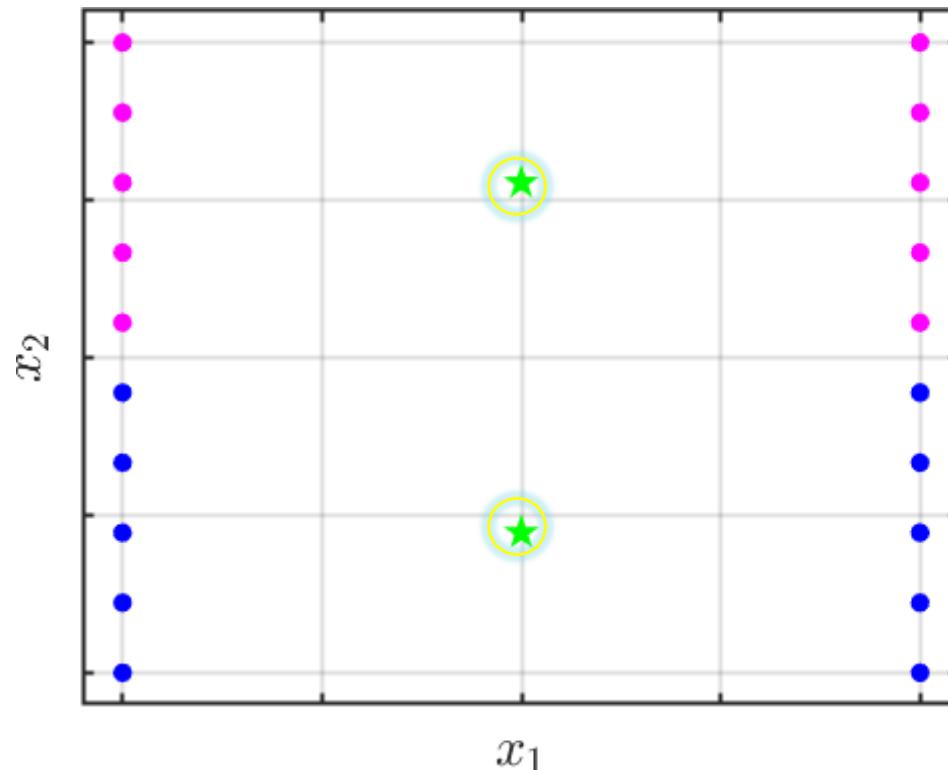
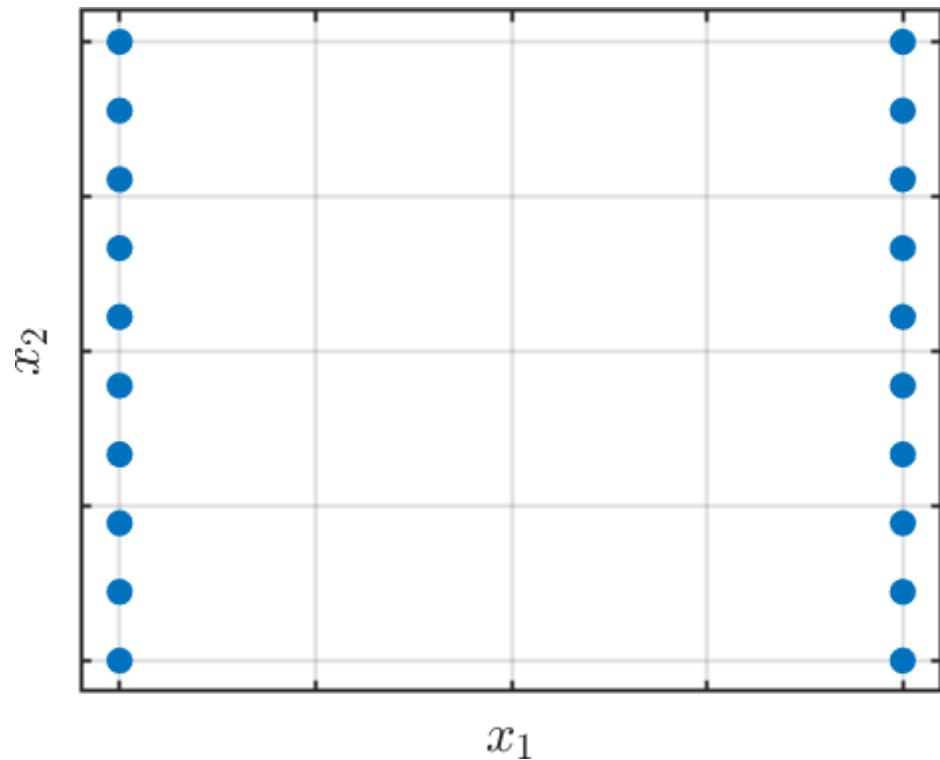
Dato	A	B
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

	A	B
C1	1.25	1.5
C2	3.9	5.1



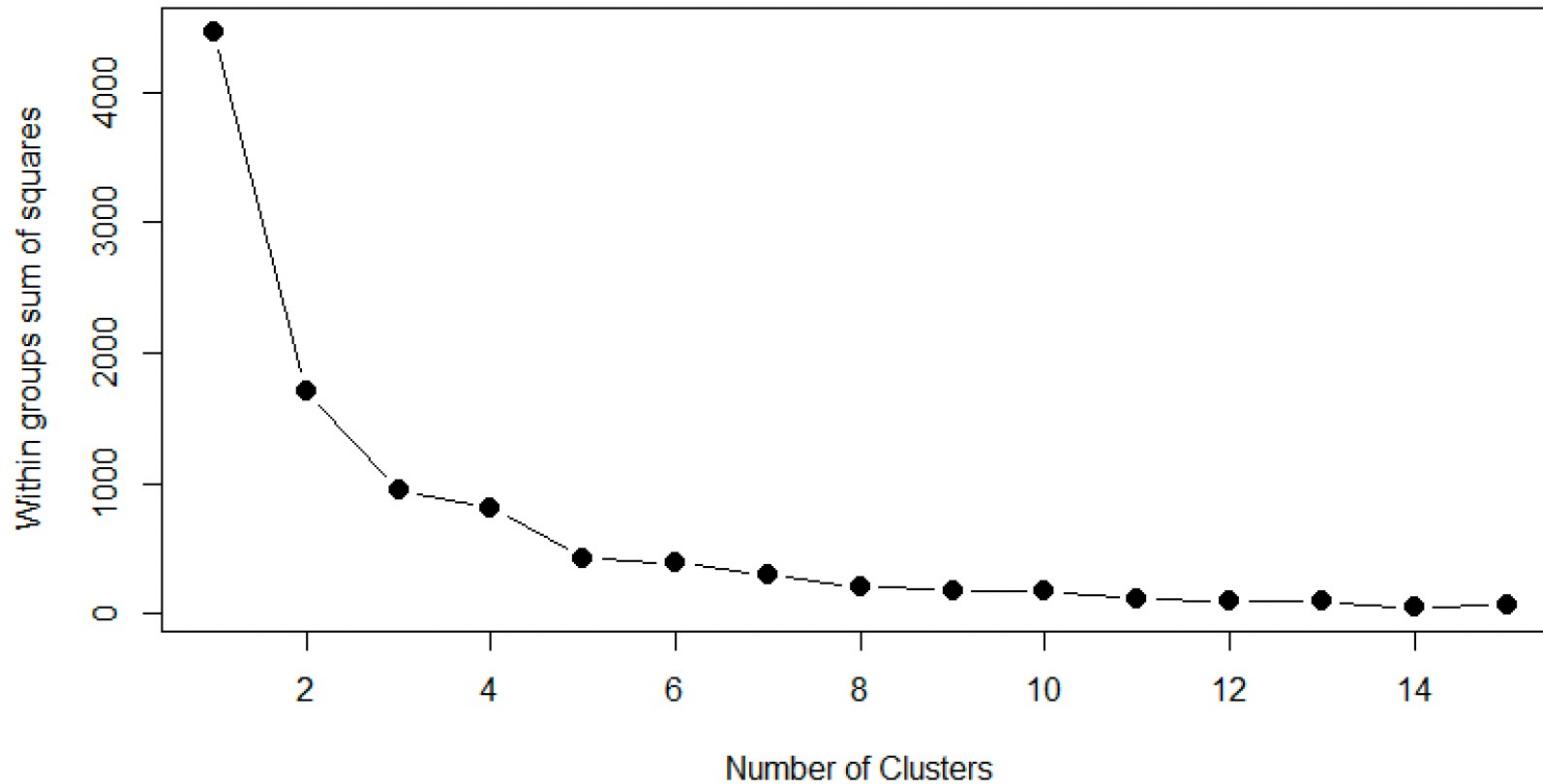
Problemas de K-Means

Depende de la elección inicial de los centros.



Cómo seleccionar el valor de k?

Assessing the Optimal Number of Clusters with the Elbow Method



K-Medianas

K-Means - Función Objetivo:

$$\min \sum_{\text{datos}} \sum_{\text{centros}} d_2(\text{dato}_i, \text{centro}_j)$$

K-Medianas - Función Objetivo:

$$\min \sum_{\text{datos}} \sum_{\text{centros}} d_1(\text{dato}_i, \text{centro}_j)$$

K-Modas

¿Qué pasa si nuestros datos son categóricos/discretos?

	A	B	C	D
D1	5	5	2	5
D2	5	4	3	3
D3	5	4	1	3
D4	4	3	4	3
D5	5	5	3	5
D6	4	3	2	2
D7	5	3	1	4

Centro #1

Centro #2

Se suma 1 si la característica es
distinta y se suma 0 si es igual

	C1	C2		G1
D1	3	4	D1	G1
D2	0	3	D2	G1
D3	1	3	D3	G1
D4	4	0	D4	G2
D5	1	4	D5	G1
D6	4	2	D6	G2
D7	3	3	D7	G2

GRUPO 1

	A	B	C	D
D1	5	5	2	5
D2	5	4	3	3
D3	5	4	1	3
D5	5	5	3	5
Moda	5	5	3	5

GRUPO 2

	A	B	C	D
D4	4	3	4	3
D6	4	3	2	2
D7	5	3	1	4
Moda	4	3	4	4

Estos son nuestros nuevos centros!

- Re-calculamos distancias y centros.

¿Cuándo termina el proceso?

R/ Cuando no hay cambios significativos en los centros.

K-Prototypes

K- Means

Los centros de los grupos son las medias o centroides.



K- Prototypes

K- Modas

Los centros de los grupos son la moda.
(Datos categóricos)



k-means (variables numéricas) + **K-modes** (variables categóricas)

Ejemplo de Segmentación



Cantidad de créditos



Montos desembolsados



Plazo solicitado



Score crediticio



Ciudad



Egreso Mensual



Ingreso mensual



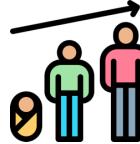
Actividad laboral



Tipo de trabajador



Nivel de estudios



Edad



Género



Estado civil



hijos



personas dependientes

Ejemplo de Segmentación

Grupo 1:

- 3 créditos por cliente
- Monto promedio **\$246.000 COP**
- Plazo promedio de **29 días**
- Edad promedio de **41 años**
- Nivel de educación superior
- Egresos cerca a **\$1'870.000 COP**
- Ingresos cerca a **\$4'270.000 COP**
- Tendencia a personas casadas y divorciadas
- Score crediticio cerca a **738**

Grupo 2:

- 2 créditos por cliente
- Monto promedio **\$204.000 COP**
- Plazo promedio de **30 días**
- Edad promedio de **34 años**
- Nivel de educación elemental (básico y bachillerato)
- Egresos cerca a **\$920.000 COP**
- Ingresos cerca a **\$1'870.000 COP**
- Tendencia a personas solteras o en unión libre
- Score crediticio cerca a **675**

Grupo 3:

- 2.5 créditos por cliente
- Monto promedio **\$195.000 COP**
- Plazo promedio de **16 días**
- Edad promedio de **32 años**
- Nivel de educación medio-alto
- Egresos cerca a **\$885.000 COP**
- Ingresos cerca a **\$2'190.000 COP**
- Tendencia a personas solteras o en unión libre
- Score crediticio cerca a **684**

Ejemplo de Segmentación

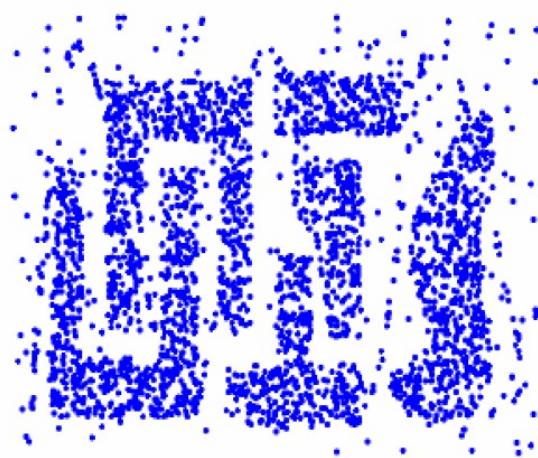
		Not Engaged	Totally Engaged	Engaged and not Engaged (50-50)	Engaged and not Engaged (Tend to Engage)	Engaged and not Engaged (Tend to NoEngage)
N Size		179	459	62	151	74
EDUCATION- highest degree	HS or less or trade degree	3%	4%	17%	25%	3%
	Some college or 2 year	13%	12%	10%	33%	10%
	4 year college degree	27%	27%	19%	16%	31%
	Post grad- some or degree	57%	58%	53%	24%	55%
EDUCATION_3way	Less than 4 yr college	16%	15%	27%	58%	13%
	4 year college	27%	27%	19%	16%	31%
	Post grad- some or degree	57%	58%	53%	24%	55%
GENDER- What is your gender?	Female	42%	57%	55%	53%	50%
	Male	58%	43%	45%	47%	50%
Age - 6way and NA	18-25	2%	0%	0%	1%	5%
	26-34	34%	18%	16%	24%	36%
	35-54	25%	39%	49%	44%	32%
	55-64	24%	17%	11%	17%	17%
	65-74	10%	18%	19%	12%	9%
	75+	5%	9%	5%	3%	1%
	Not answered	0%	0%	0%	0%	0%

Ejemplo de Segmentación

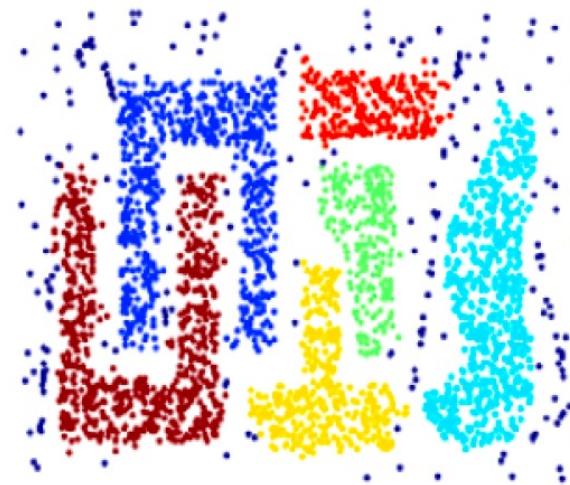
Not Engaged	Totally Engaged	Engaged and not Engaged (50-50)	Engaged and not Engaged (Tend to Engage)	Engaged and not Engaged (Tend to NoEngage)
High level of education	High level of education	High level of education	Tend to less than 4 year college	High level of education
Male	Female	Female	50-50 gender	50-50 gender
Younger people	Middle Age	Middle Age	Middle Age	Younger people
Mostly Asia, Indian and white	mostly white people	Mostly Asia, Indian, hispanic and white	most presence of hispanic, black and asian in comparison to other groups	White people and most presence of multiracial in comparison to other groups
does not have a predominant frequency on getting communications	read/hear/watch info/news of the county	read/hear/watch info/news of the county	does not have a predominant frequency on getting communications	Tend to not watch/hear/read news
Frequent use of mobile phone and tablet	Frequent use of mobile phone and tablet, as well as printed newsletter and paperletter	Frequent use of mobile phone and tablet, as well as printed newsletter and paperletter	Frequent use of mobile phone and live TV	Tend to not use any source to get information
Does not consider important comunications in other language than english	Does not consider important communications in other language than english	Considers relevant get information in other language than english	Considers relevant get information in other language than english	Does not consider important comunications in other language than english

Agrupación basado en densidad

Los algoritmos basados en densidad, tratan de formar agrupaciones en áreas con altas densidades de ejemplos



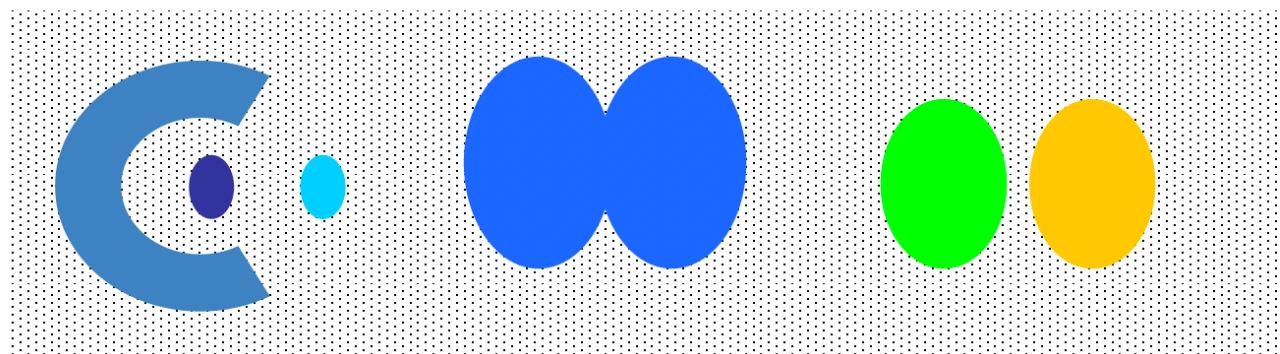
a) Datos originales



b) Datos después de clustering

Agrupación basado en densidad

- Un cluster en una región densa de puntos, separada por regiones poco densas de otras regiones densas.
- Útiles cuando los clusters tienen formas irregulares, están entrelazados o hay ruido/outliers en los datos.



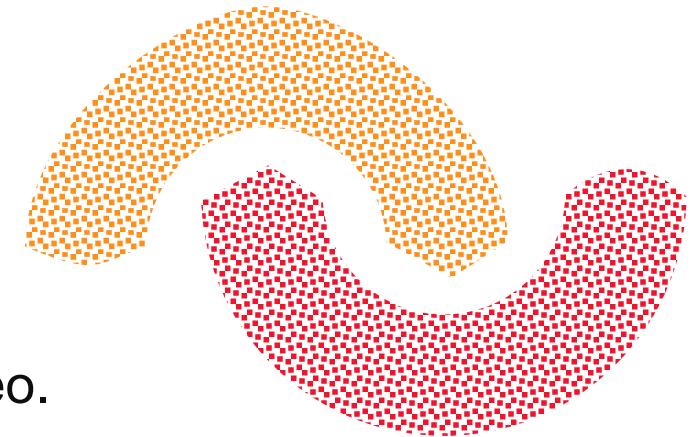
Agrupación basado en densidad

Ventajas:

- ✓ Identifica clusters de formas arbitrarias (no necesariamente convexos)
- ✓ Robusto ante la presencia de ruido
- ✓ Escalable sin importar el conjunto de datos

DBSCAN

Density Based Spatial Clustering of Application with Noise



1. Para cada dato: ¿Cuántos datos están a distancia ϵ ?
2. Si un dato tiene k datos en su ϵ -vecindario entonces es el núcleo.
3. Los datos cercanos a un núcleo pertenecen al mismo grupo.

✓ **Anomalía:**

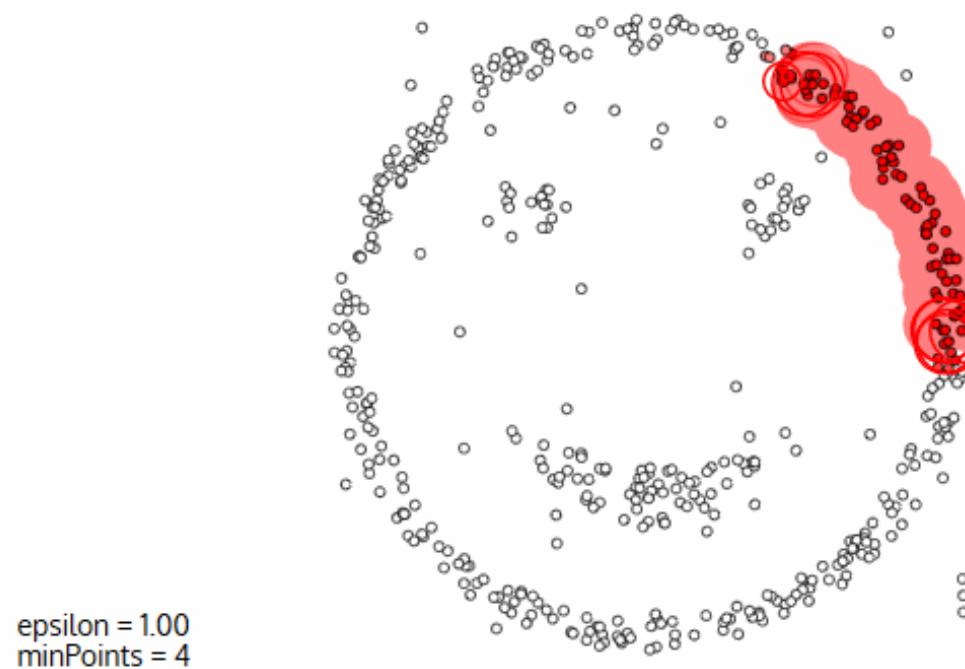
- ✓ Datos que no son núcleo y que no pertenecen a ningún grupo.

✓ **Ventaja:**

- ✓ Este algoritmo funciona bien si todos los clúster son densos y están bien separados por regiones de baja densidad.

DBSCAN

Es un algoritmo muy sencillo que puede identificar grupos que tienen diferentes formas (No solo circulares y/o rectangulares)



Restart



Pause

DBSCAN

Ventajas:

- ✓ El número de clusters no es un parámetro.
- ✓ Formas geométricas arbitrarias.
- ✓ Robusto detectando outliers.
- ✓ No es susceptible al orden de los datos en la base de datos.

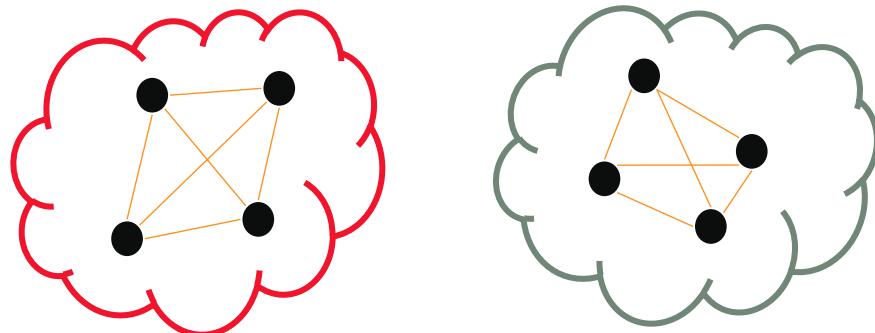
Desventajas:

- ✗ No es enteramente determinista.
- ✗ Depende de la distancia que se use.
- ✗ Problemático cuando los grupos tienen densidades muy distintas.

Métricas de evaluación

Cohesión

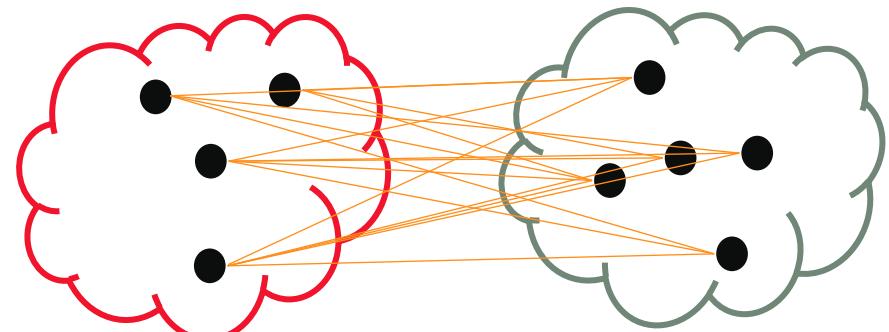
Distancia entre los elementos de un grupo.



$$\sum_i \sum_{(x,y) \in G_i} d(x, y)$$

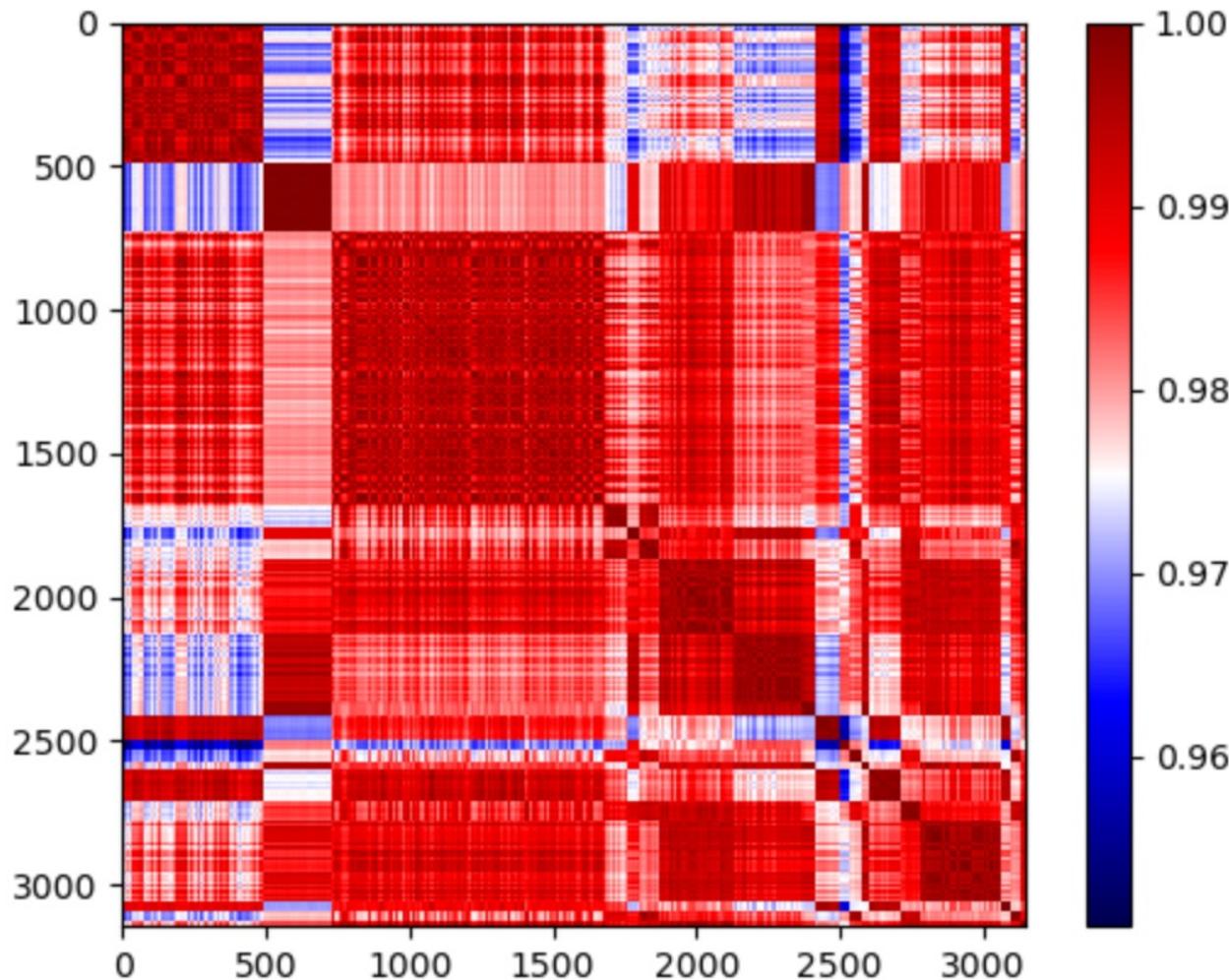
Separación

Distancia entre los grupos



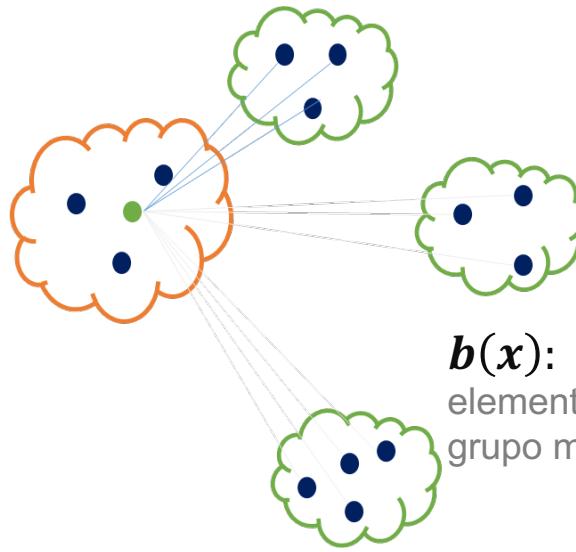
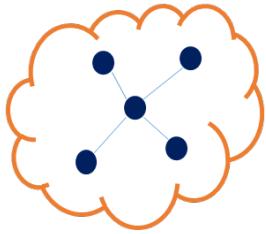
$$\sum_{x \in G_i} \sum_{\substack{y \in G_j \\ G_j \neq G_i}} d(x, y)$$

Matriz de Similitud



Métricas de desempeño de un cluster: Coeficiente de silueta

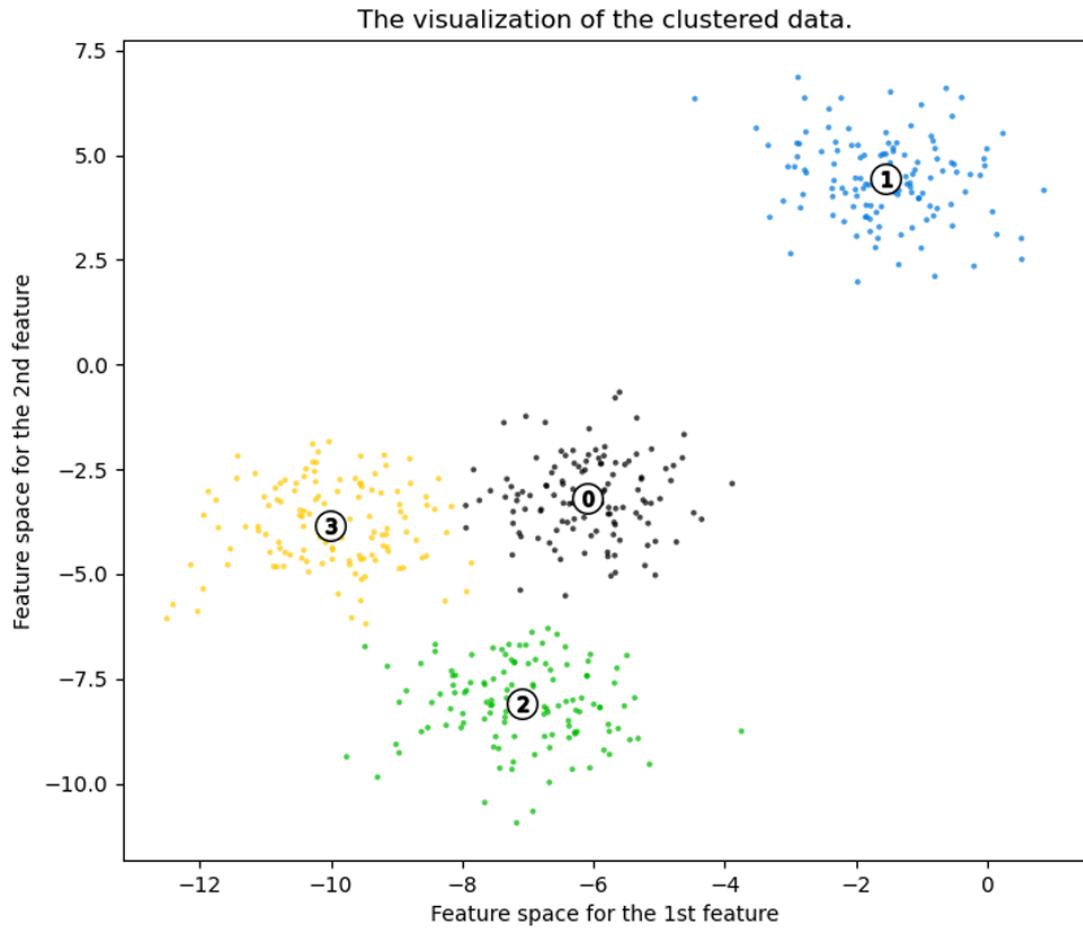
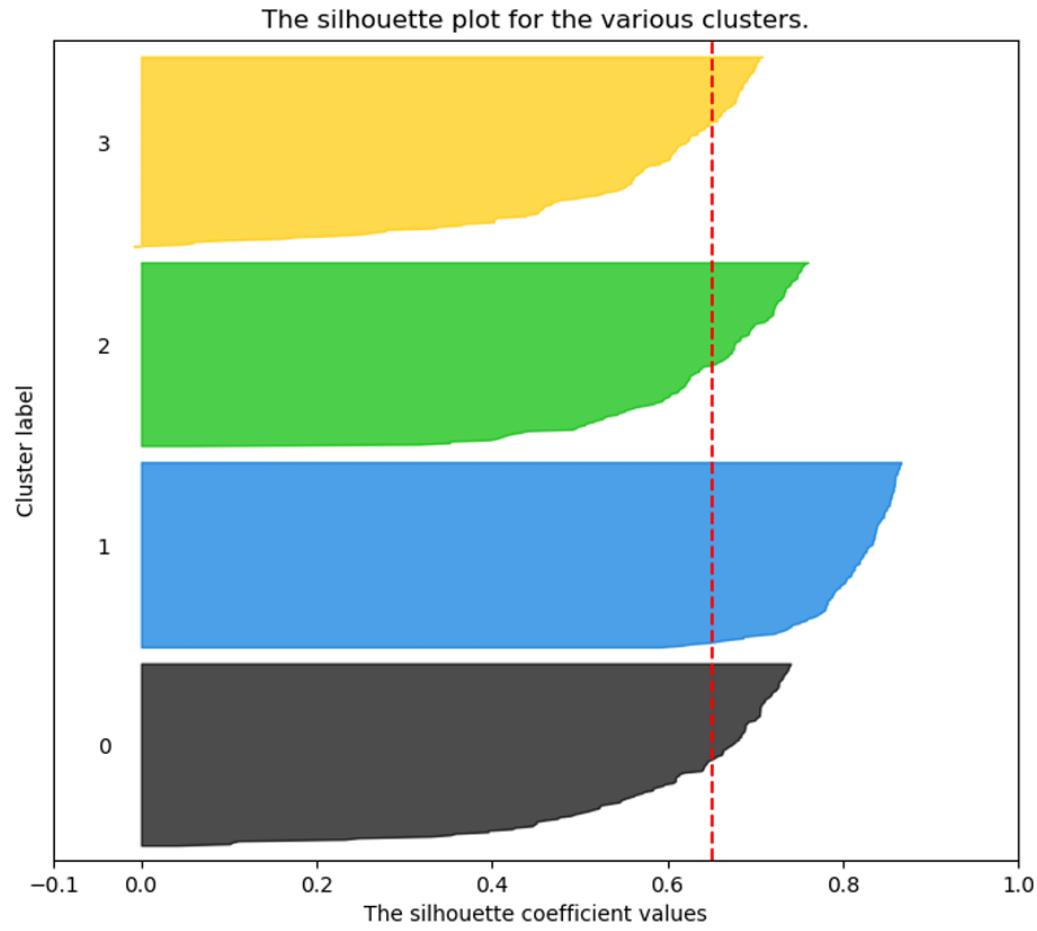
$a(x)$: Distancia promedio de un elemento a los demás elementos del grupo.



$b(x)$: Distancia promedio de un elemento a todos los elementos del grupo mas cercano.

$$SC = \frac{1}{N} \sum_{i=1}^N \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}$$

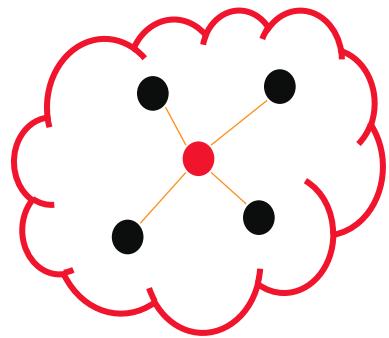
Métricas de desempeño de un cluster: Coeficiente de silueta



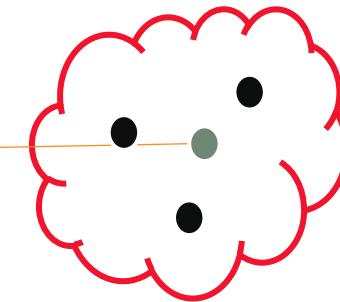
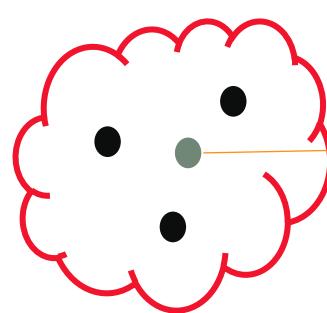
Davies-Bouldin Index

“Similitud” promedio entre los grupos.

Los valores más cercanos a cero indican una mejor partición.



Si: Distancia promedio de los elementos de un grupo a su centroide.



r_{ij}: Distancia entre dos centroides

$$DB = \frac{1}{K} \sum_{i=1} \max_{i \neq j} \frac{s_i + s_j}{r_{ij}}$$

DBCV

Moulavi, D., Jaskowiak, P. A., Campello, R. J., Zimek, A., & Sander, J. (2014, April). Density-based clustering validation. In *Proceedings of the 2014 SIAM international conference on data mining* (pp. 839-847). Society for Industrial and Applied Mathematics.

**Los valores de este índice varían entre -1 y 1,
dónde 1 es la mejor partición**

**Inspira
Crea
Transforma**