Técnicas de clasificación de individuos (I): Análisis Clúster Jerarquizado.

Miguel Ángel Tarancón Morán & Consolación Quintana Rojo

Área de Estadística Económica y Empresarial

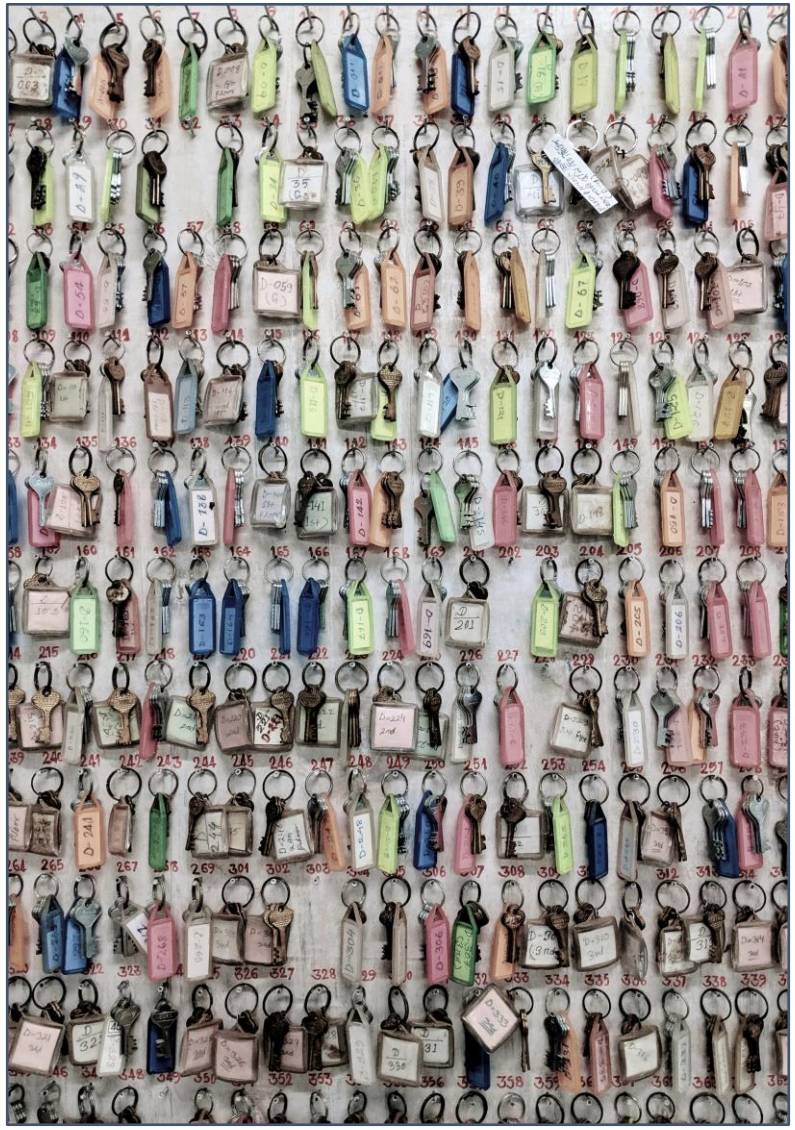
Departamento de Economía Aplicada 1

Universidad de Castilla – La Mancha

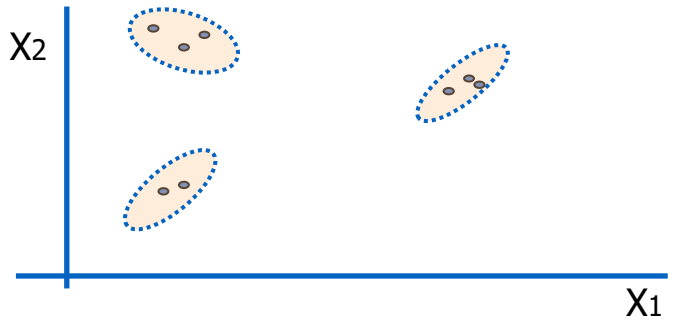
Contenidos.

1. Introducción.
2. Distancia entre casos.
3. Métodos de agrupación jerárquicos.
4. Caracterización de grupos.

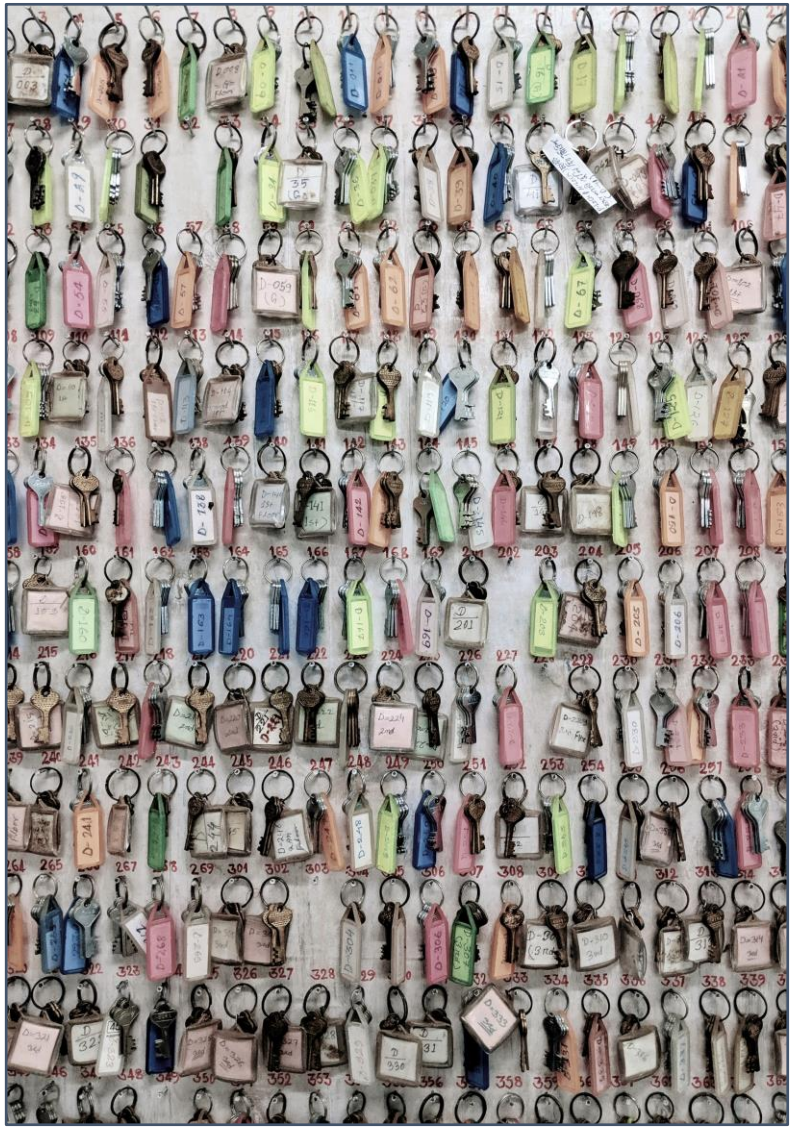
Introducción.



- El análisis de conglomerados o análisis clúster (AC) trata de clasificar individuos o casos en grupos de manera que:
  - Cada grupo, conglomerado o clúster contenga a los **casos más parecidos** entre sí, en términos de una serie de variables (variables clasificadoras).
  - Los **grupos** sean lo más **distintos** posible entre sí, de acuerdo con las variables consideradas.



Introducción.



- **Proceso** del AC para la agrupación de individuos:
  - Se cuenta con **n** casos, y para cada uno tenemos el valor de **m** variables clasificadoras.
  - Se establece una **medida de proximidad** que cuantifica lo que dos casos o grupos de individuos se parecen, en función de los valores que presentan en cada variable clasificadora.
  - Se crean los **grupos** con las observaciones que entre sí tengan una **menor distancia**. Existen dos tipos de esquemas para crear los grupos: **jerárquicos** y **no-jerárquicos**.
  - **Se describen los grupos obtenidos** y se comparan unos con otros.

Distancia entre casos.

- El primer paso en el AC es **cuantificar lo próximo** que están los individuos o casos (o grupos previamente creados) teniendo en cuenta las variables clasificadoras; ya que la idea es formar grupos de individuos “muy próximos” entre sí, siendo cada grupo, a su vez, muy distante del resto.

- Las medidas de proximidad más utilizadas son la **distancia euclídea** y la **distancia euclídea al cuadrado**:

$$d_{ii'} = \sqrt{\sum_j \left(x_{ij} - x_{i'j}\right)^2}$$ ← Distancia euclídea entre los individuos i e i’

$$d^2_{ii'} = \sum_j \left(x_{ij} - x_{i'j}\right)^2$$ ← Distancia euclídea al cuadrado entre los individuos i e i’

- Estas medidas son muy sensibles a la escala de las variables. Para eliminar este efecto distorsionador, se trabaja con las **variables tipificadas**.

Distancia entre casos.

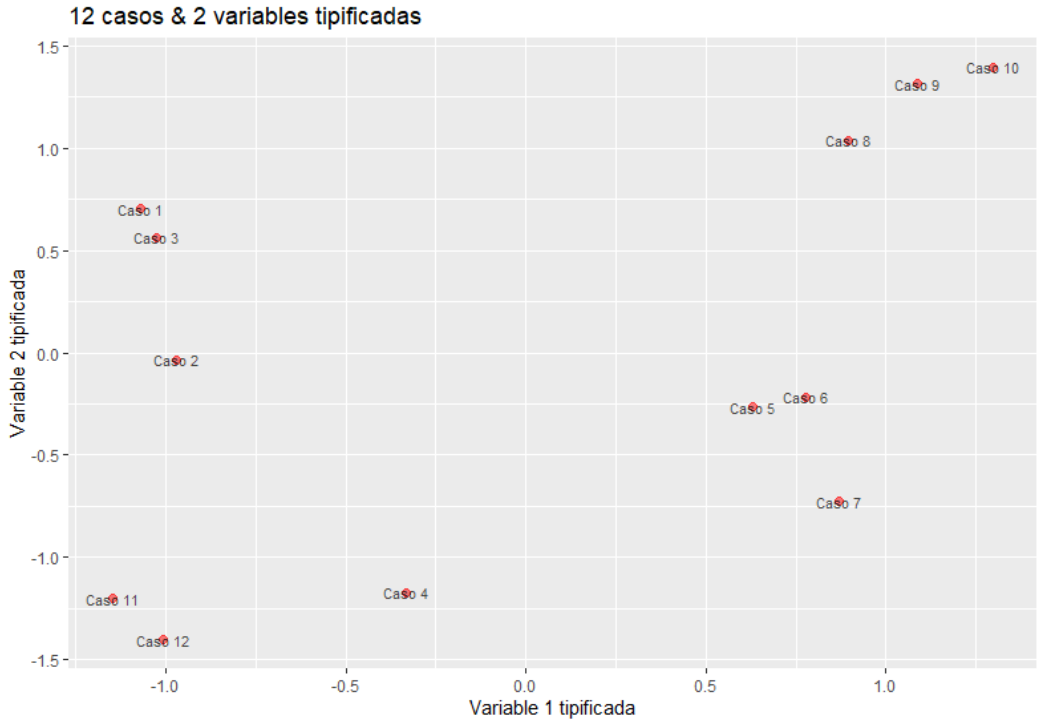
- **Ejemplo:** 12 casos y 2 variables clasificadoras. Las variables deben ser previamente tipificadas.

Casos	variable.1	variable.2
Caso 1	9,32	431,00
Caso 2	11,20	322,00
Caso 3	10,15	410,00
Caso 4	23,60	154,00
Caso 5	42,30	288,00
Caso 6	45,20	295,00
Caso 7	47,00	220,00
Caso 8	47,50	480,00
Caso 9	51,20	521,00
Caso 10	55,30	533,00
Caso 11	7,80	150,00
Caso 12	10,50	120,00
Media	30,0891667	327
Desv. típica	19,44	147,50

Casos	variable.1	variable.2
Caso 1	-1,07	0,71
Caso 2	-0,97	-0,03
Caso 3	-1,03	0,56
Caso 4	-0,33	-1,17
Caso 5	0,63	-0,26
Caso 6	0,78	-0,22
Caso 7	0,87	-0,73
Caso 8	0,90	1,04
Caso 9	1,09	1,32
Caso 10	1,30	1,40
Caso 11	-1,15	-1,20
Caso 12	-1,01	-1,40
Media	0,00	0,00
Desv. típica	1,00	1,00

Distancia entre casos.

- **Ejemplo:** Ahora se calculan las distancias entre los casos, según las variables clasificadoras (originales tipificadas).



- Los casos se agruparán según el valor de las distancias (euclídeas) que tienen respecto al resto de casos.

Métodos de agrupación jerárquicos.

- Hay dos tipos de métodos de agrupación de casos: **jerárquicos** y **no-jerárquicos**.
- Es importante seleccionar un método adecuado, ya que pueden aportar **soluciones muy distintas**.
- En los **métodos jerárquicos,** se van formando sucesivamente grupos como agrupación de otros grupos precedentes, hasta llegar a un único grupo que recoge a todos los individuos; tomando el proceso una **estructura piramidal**.

Métodos de agrupación jerárquicos.

- ¿Cuándo utilizar los métodos jerárquicos?
  - Cuando hay **pocos casos** que clasificar.
  - Cuando nuestro objetivo pasa por crear **grupos que recojan a todos los casos**, más que definir simplemente tipologías más o menos homogéneas de casos (lo que se obtiene caracterizando los grupos obtenidos), incluidos los *outliers*. Ejemplo: queremos repartir en grupos homogéneos a los países de la UE de acuerdo con varias variables económicas clasificadoras.
  - Cuando **se desconoce el número de grupos** a formar. No obstante, existen algoritmos que sugieren un número de grupos a crear.

Métodos de agrupación jerárquicos.

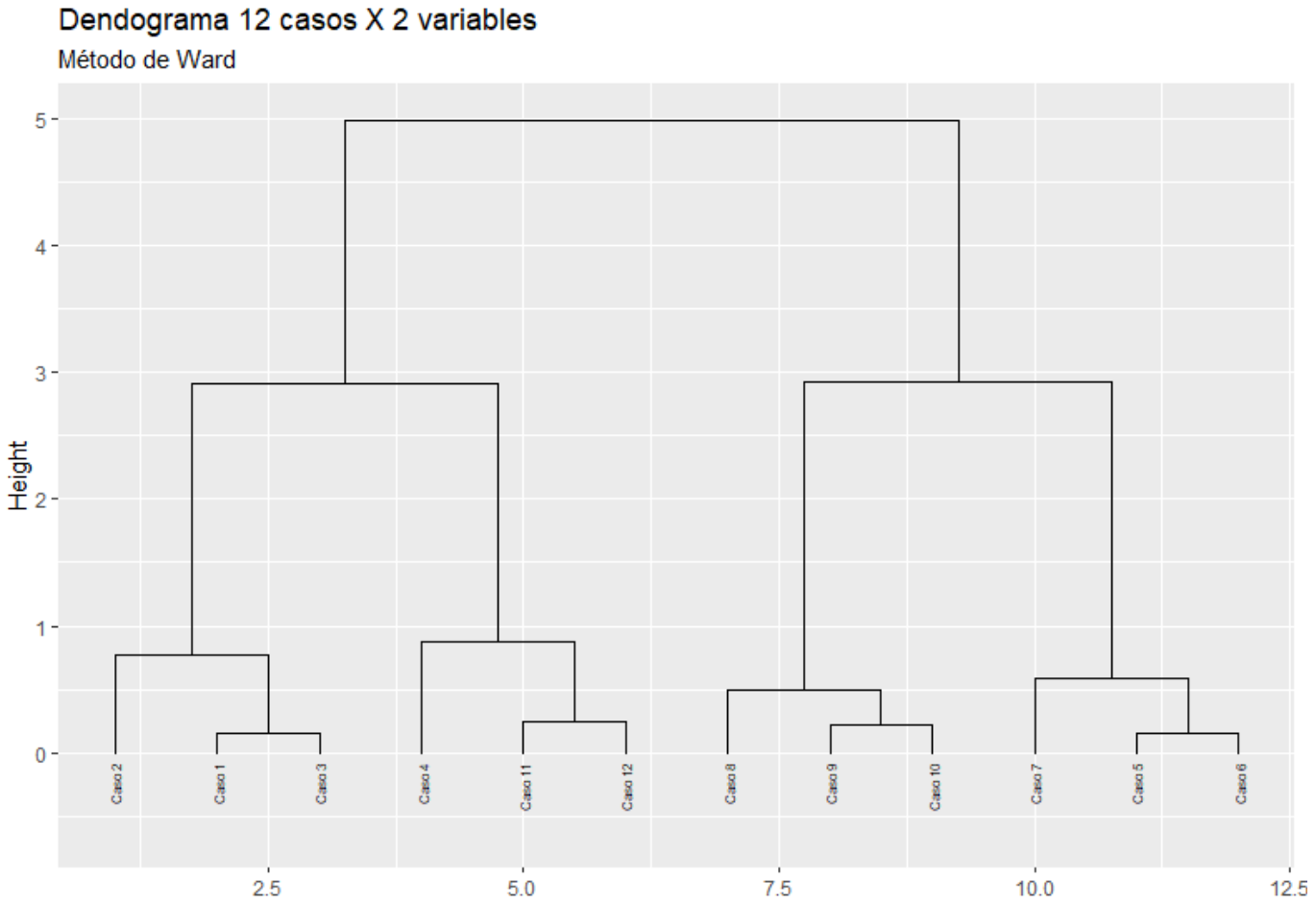
- ¿Qué métodos jerárquicos existen?
  - **Método del vecino más cercano (single linkage):** la distancia que se considera entre grupos es la distancia entre sus elementos más próximos.
  - **Método del vecino más lejano (complete linkage):** la distancia que se considera entre grupos es la distancia entre sus elementos más lejanos.
  - **Método de Ward (Ward method):** se unen los grupos que dan lugar a otro grupo cuyos casos tienen una menor suma de los cuadrados de sus distancias respecto al centro de dicho grupo (menor varianza).
  - **Otros métodos:** vinculación intergrupos (average linkage between groups), vinculación intragrupos (whithin group)…

Métodos de agrupación jerárquicos.

- ¿Qué método jerárquico elegir?
  - Cada método proporciona soluciones que pueden variar mucho. Se puede probar con varios métodos y se seleccionará la **solución que parezca más coherente** desde el punto de vista teórico; **y estable** desde el punto de vista empírico.
  - En la práctica, uno de los métodos más utilizados es el de **Ward,** porque proporciona grupos muy homogéneos.

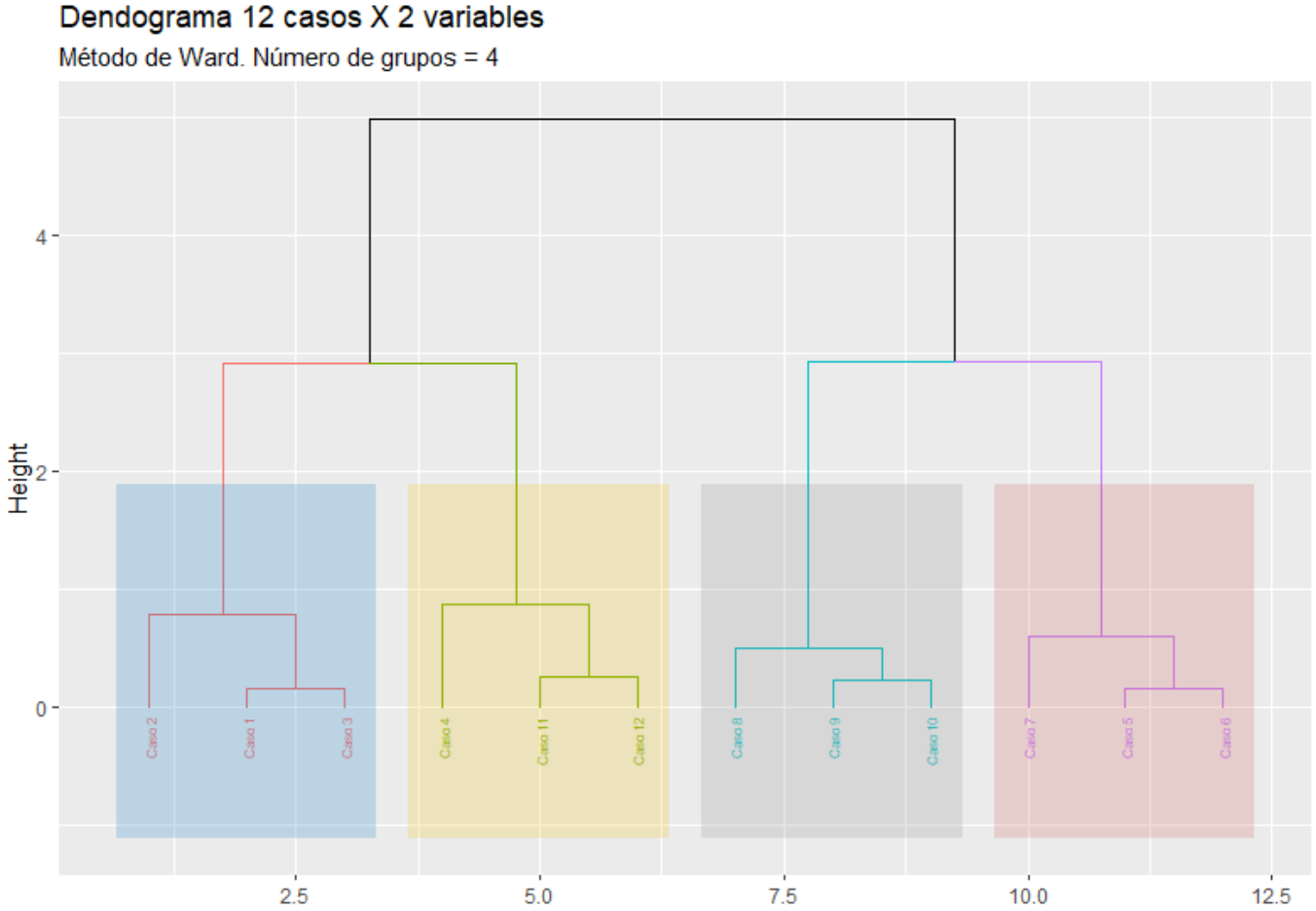
Métodos de agrupación jerárquicos.

- **Ejemplo:** método de Ward. Resultado (dendograma).



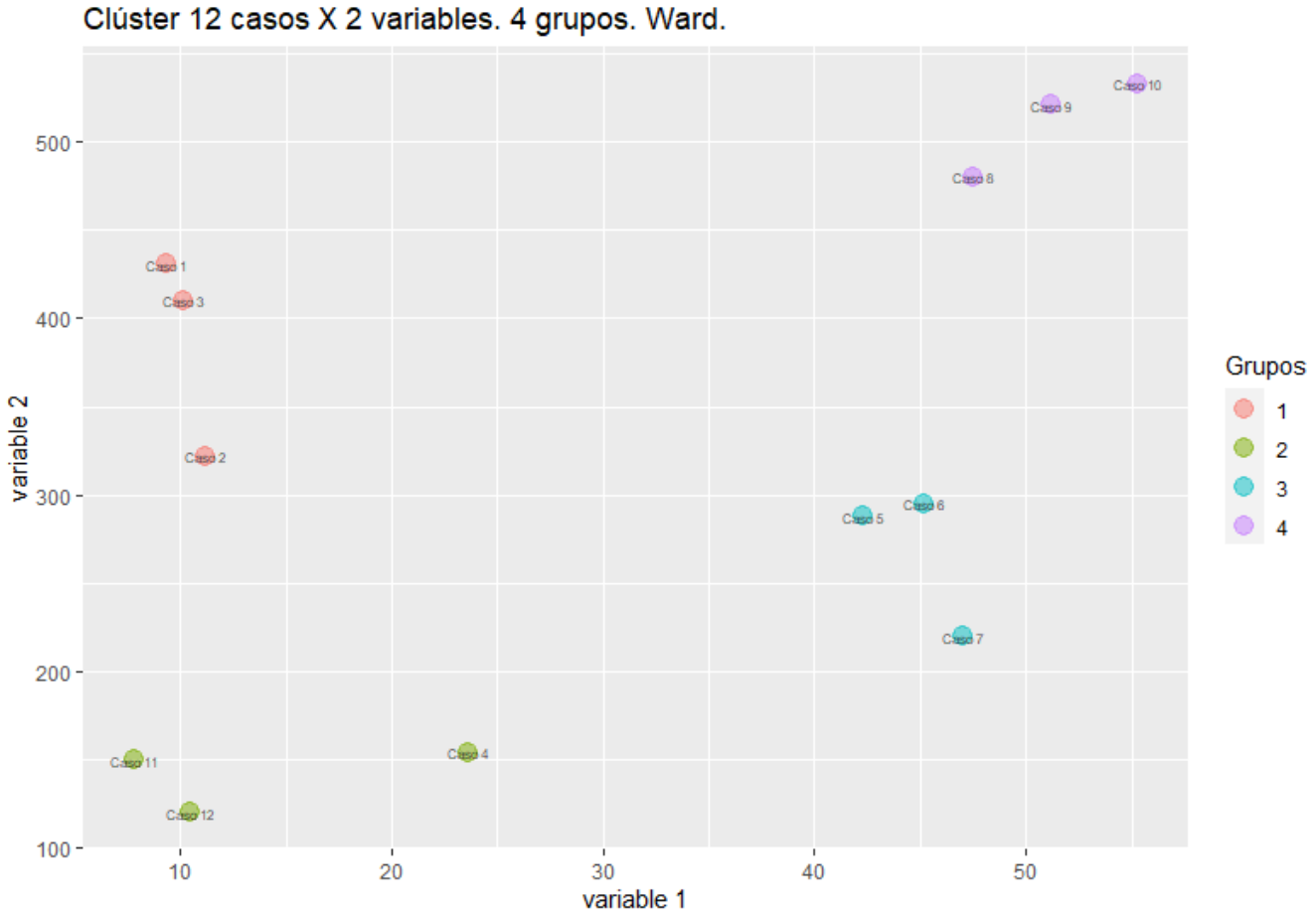
Métodos de agrupación jerárquicos.

- **Ejemplo:** ¿Cómo decidimos el número de grupos? Si no se tiene una idea propia, hay algoritmos como **NbClust** (R).

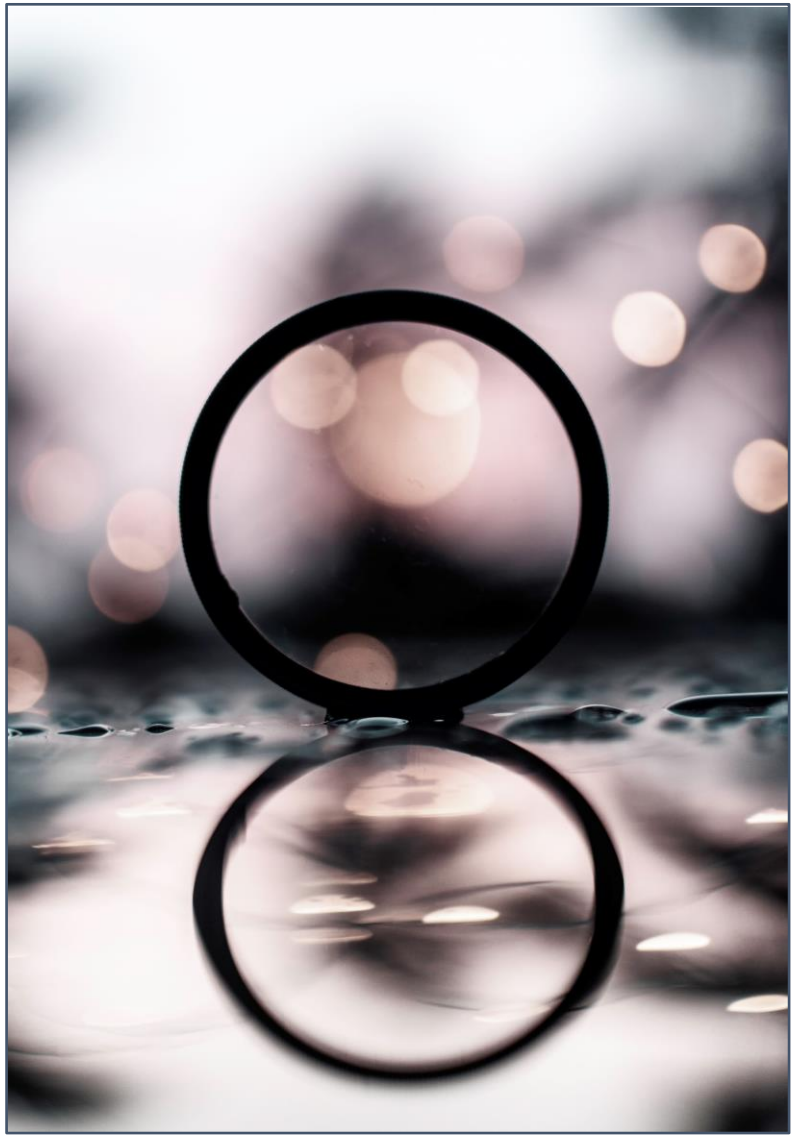


Métodos de agrupación jerárquicos.

- **Ejemplo:** En un diagrama de dispersión con las dos variables:



Caracterización de grupos.



- Una vez establecidos los conglomerados, clústeres o grupos; es importante caracterizarlos para **discernir en qué se distinguen unos de los otros** principalmente.

- Una opción sencilla es establecer las **coordenadas de los centroides** de cada grupo (en función de las variables clasificadoras; pero sin tipificar), y comentar las diferencias observadas.

- Se puede hacer un **análisis de la varianza** de las variables que sirvieron para realizar el análisis clúster, para los grupos formados.

¡Muchas gracias!