Técnicas de clasificación de individuos (II): Análisis Clúster No-Jerarquizado. K-medias.

Miguel Ángel Tarancón Morán & Consolación Quintana Rojo

Área de Estadística Económica y Empresarial

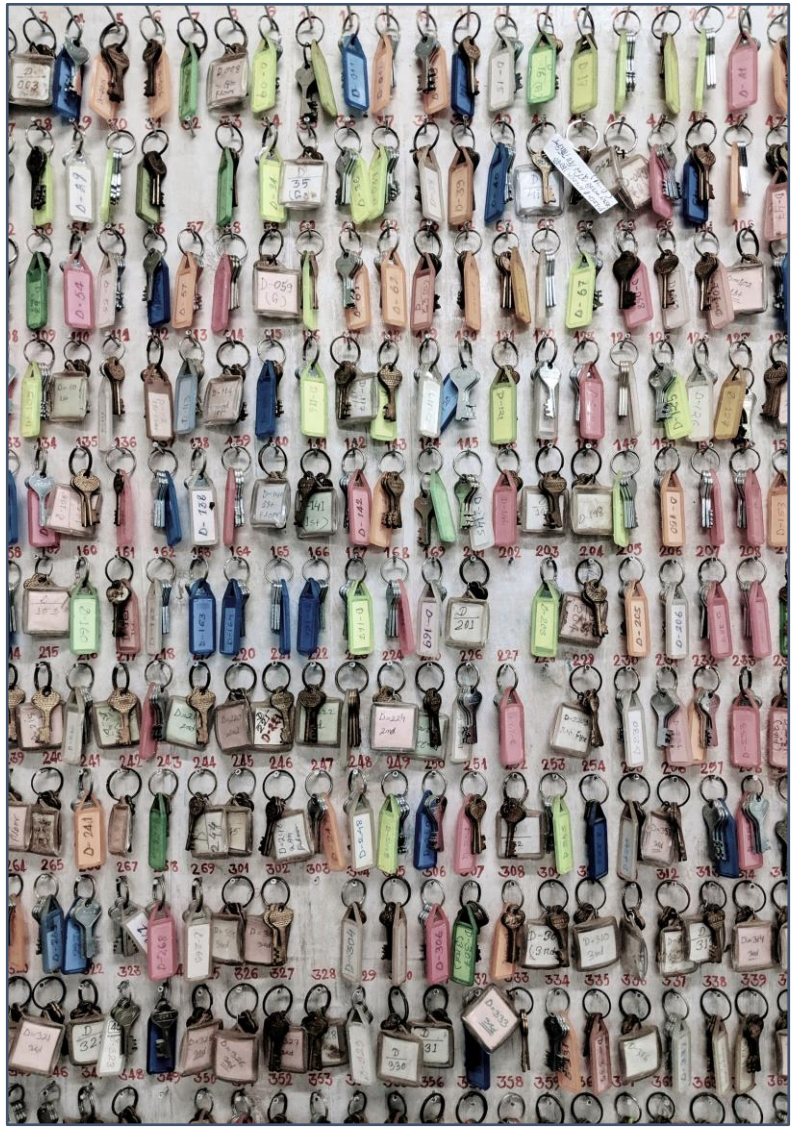
Departamento de Economía Aplicada 1

Universidad de Castilla – La Mancha

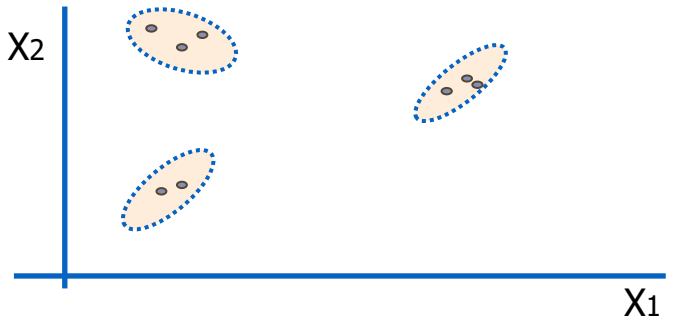
Contenidos.

1. Introducción.
2. Método de k-medias.
3. Caracterización de grupos.

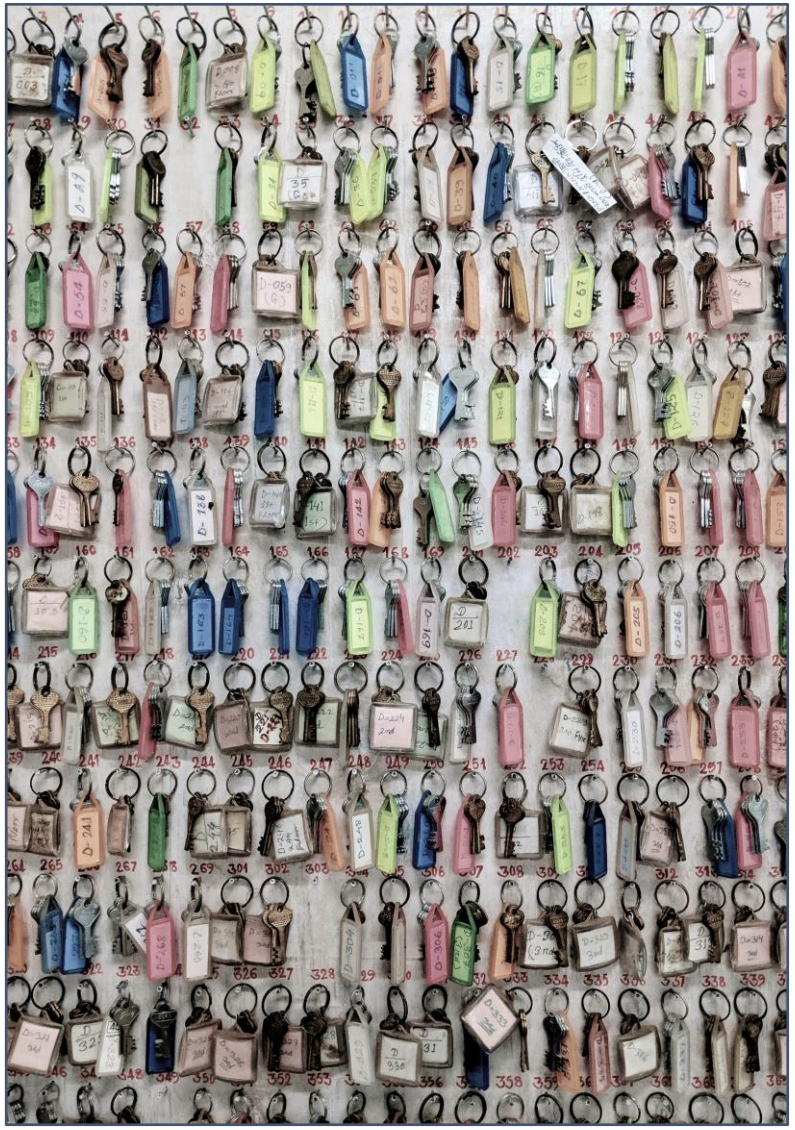
Introducción.



- **Recordatorio:** 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.

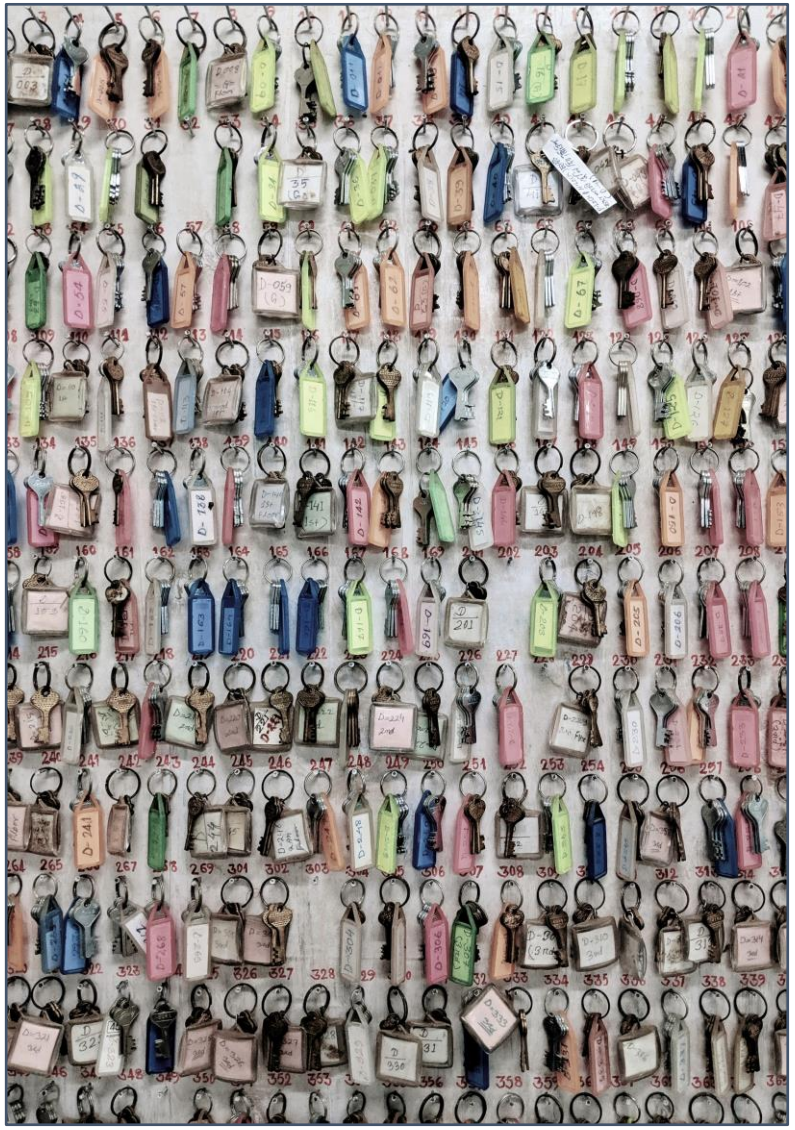


Introducción.



- **Recordatorio: proceso** del AC:
  - 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. Las más habituales: **distancia euclídea**, y distancia euclídea al cuadrado.
  - 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.

Introducción.

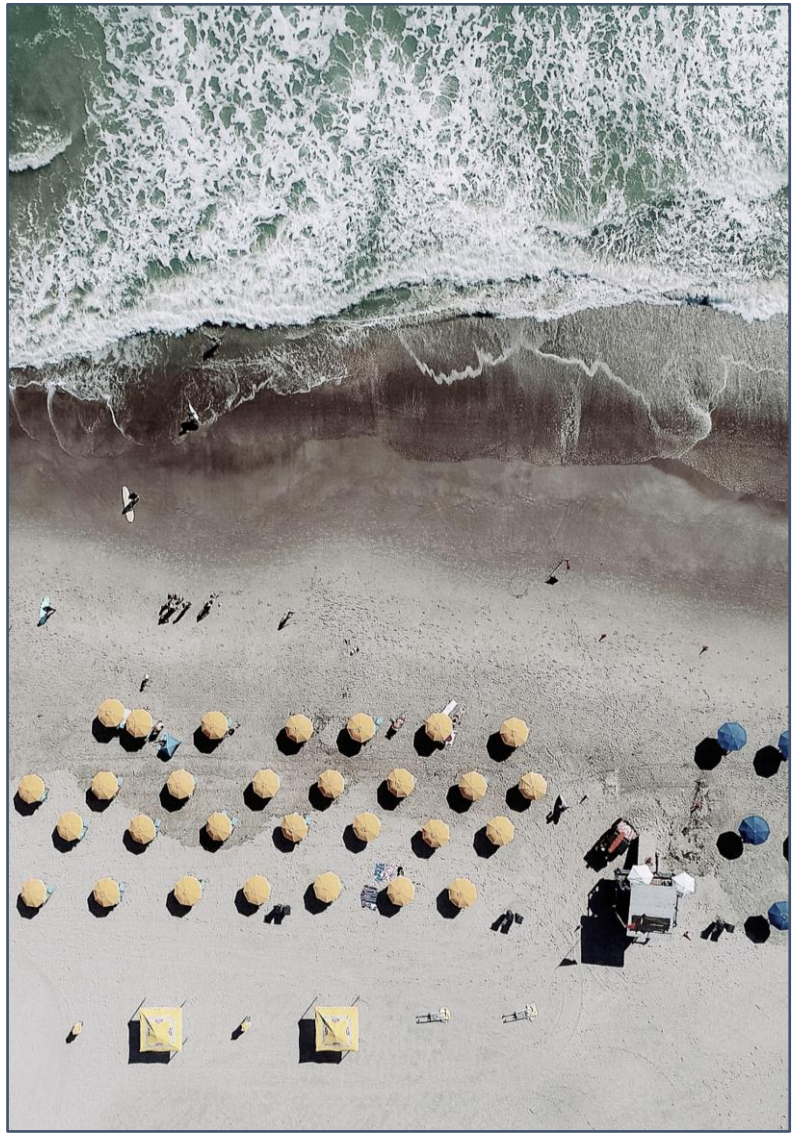


- Los **métodos de agrupación no-jerárquicos** se suelen utilizar para estudios en los que hay un **gran número de casos**.

- Se suelen utilizar cuando nuestro objetivo pasa por crear grupos que definan una **tipología de casos o individuos**, más que clasificar casos o individuos concretos. En definitiva, identificar subpoblaciones a partir de una muestra. Por eso es conveniente, previamente, detectar los *outliers* y, en su caso, **eliminarlos**, ya que podrían distorsionar las características de los grupos. Ejemplo: queremos encontrar perfiles de clientes de una gran superficie a partir de distintas variables clasificadoras, utilizando para ello una muestra de 300 clientes. A estos perfiles se les diseñarán distintas políticas comerciales.

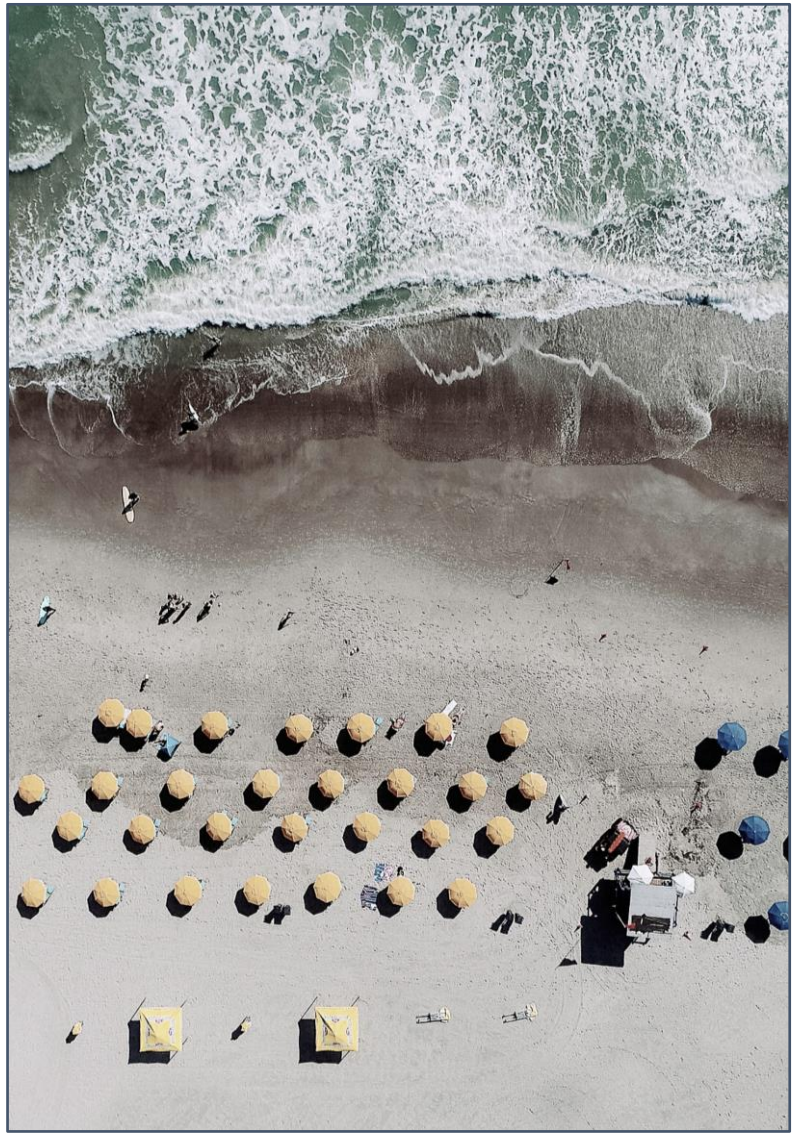
- En estos métodos, se establece **a priori** el número de grupos a formar.

Método de k-medias.



- El principal método no-jerárquico es el de **k-medias**.
- k-medias es un método **iterativo**. Se establece un **“centroide”** inicial (“*semilla*”) para cada uno de los k grupos que se quieren crear, y se van asignando a cada grupo los casos que se sitúen más cerca de su centro.
- Una vez asignados los casos, se recalculan los “centroides” de los grupos, y **se repite** el proceso en una nueva iteración.
- El procedimiento termina cuando el algoritmo encuentra la solución **convergente** (estable).
- Un ejemplo visual e interactivo se puede encontrar en: https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Método de k-medias.

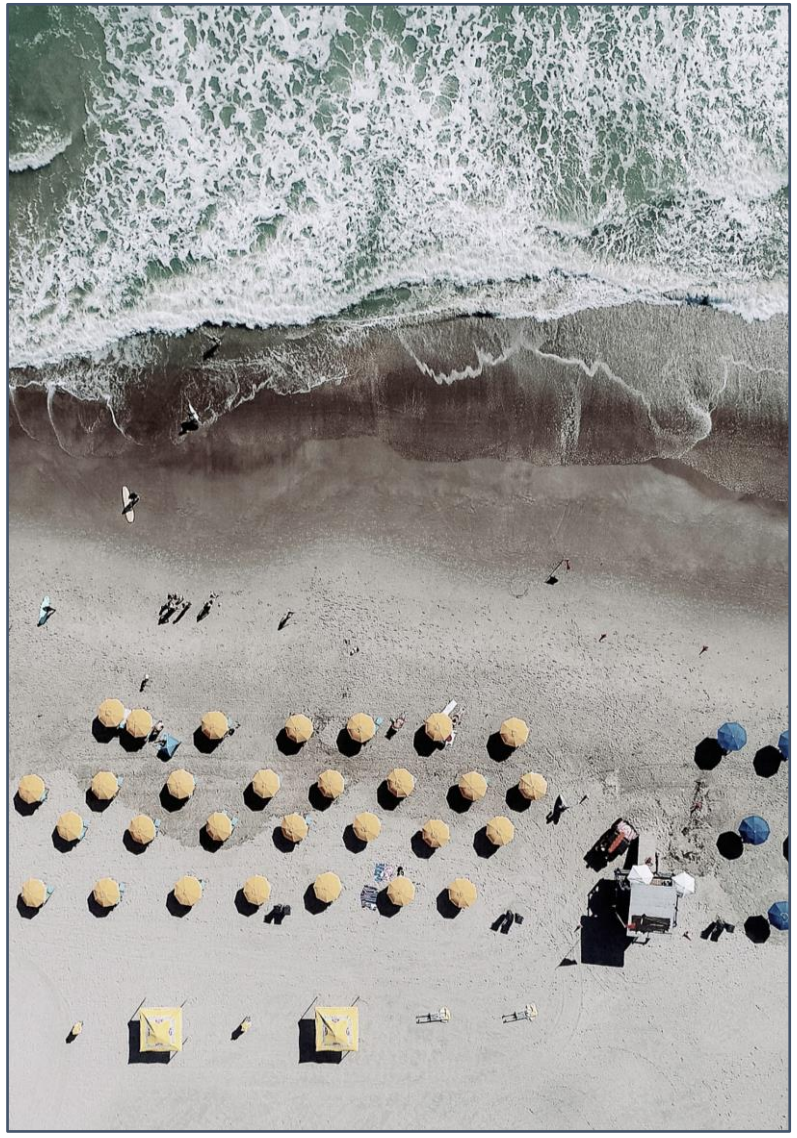


- El método de k-medias posee la ventaja, frente a los métodos jerárquicos, de que **se puede alterar la pertenencia de un caso a un grupo** en las sucesivas iteraciones, lo que lo convierte en un método, en general, más eficiente.

- Como desventaja, la solución es **muy sensible** a los **“centroides”** (“semillas”) propuestos al inicio del proceso.

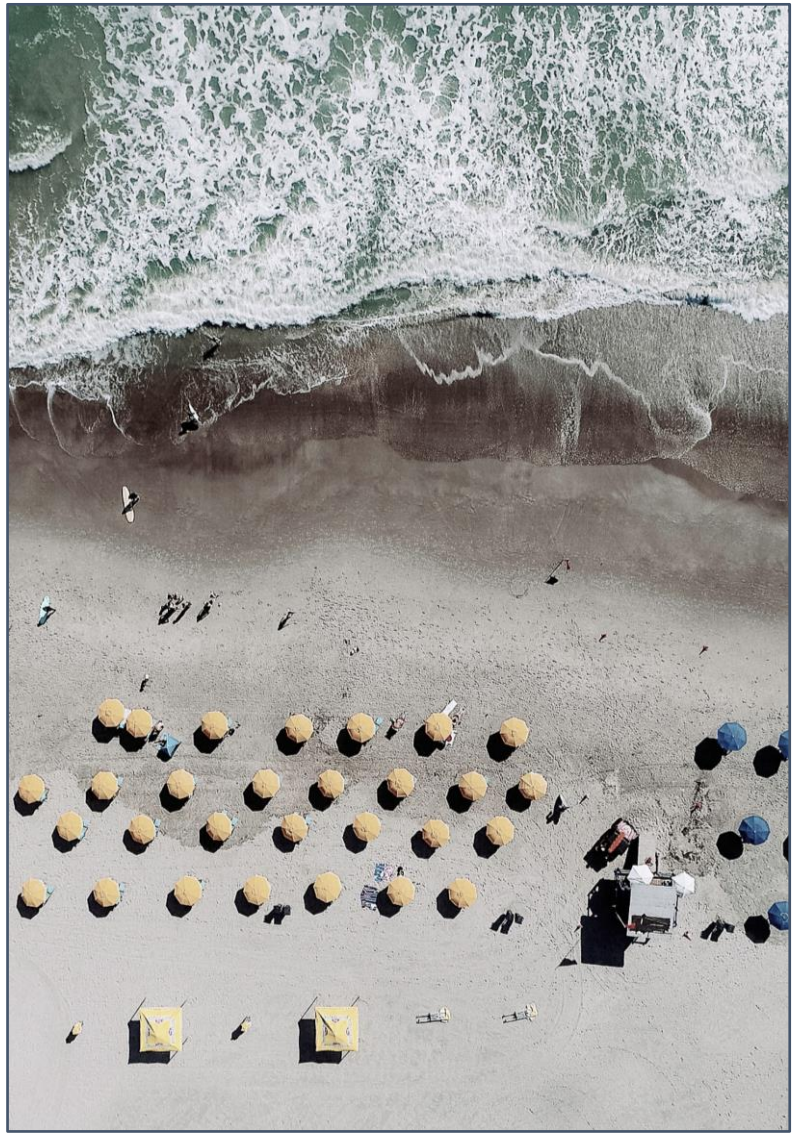
- Además, hay que partir de un **número de grupos establecido a priori**.

Método de k-medias.



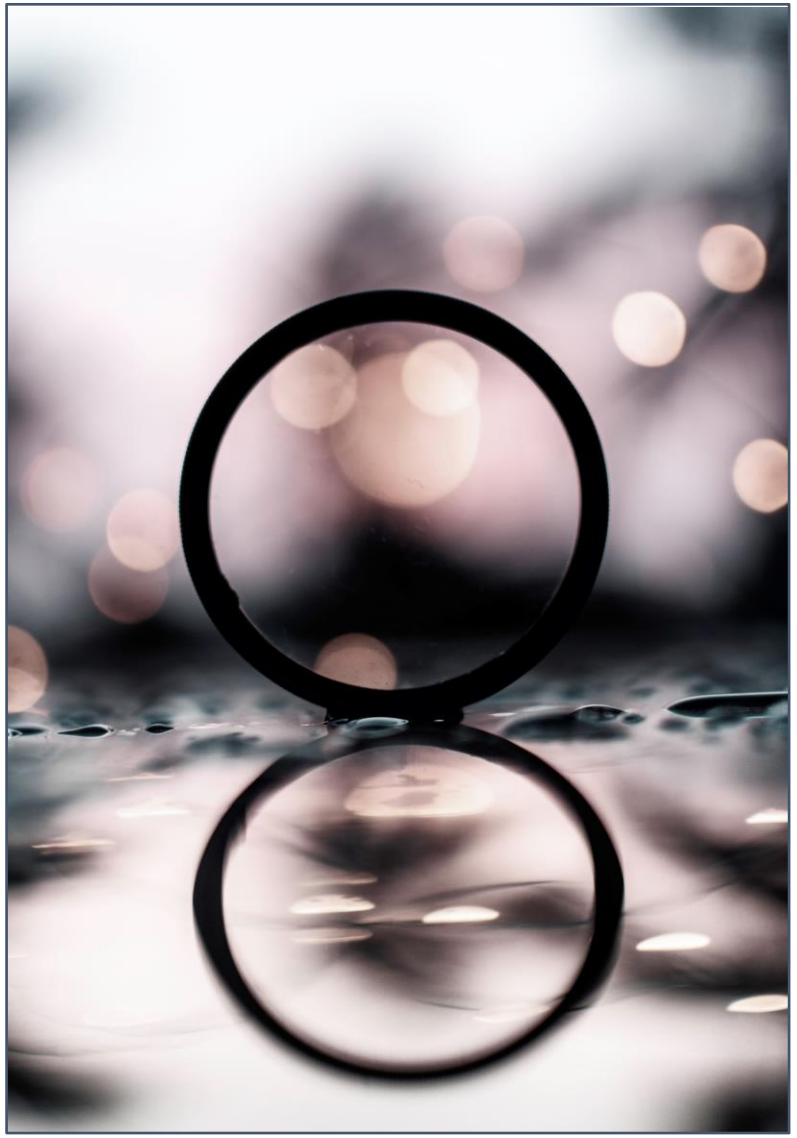
- ¿Cómo fijar un **número de grupos** a priori?
- Hay ocasiones en que el investigador establecerá un número que le sea manejable o útil según los objetivos que persiga.
- Cuando no se tenga un número claro de grupos o conglomerados a formar, se podrá optar por probar con varios números, y evaluar las soluciones obtenidas.
- También puede ayudar el realizar previamente un análisis jerárquico para estudiar el dendograma.
- Además, existen algoritmos, como **NbClust** (R), que sugieren un número de grupos en función de una batería de pruebas presentes en la literatura.

Método de k-medias.



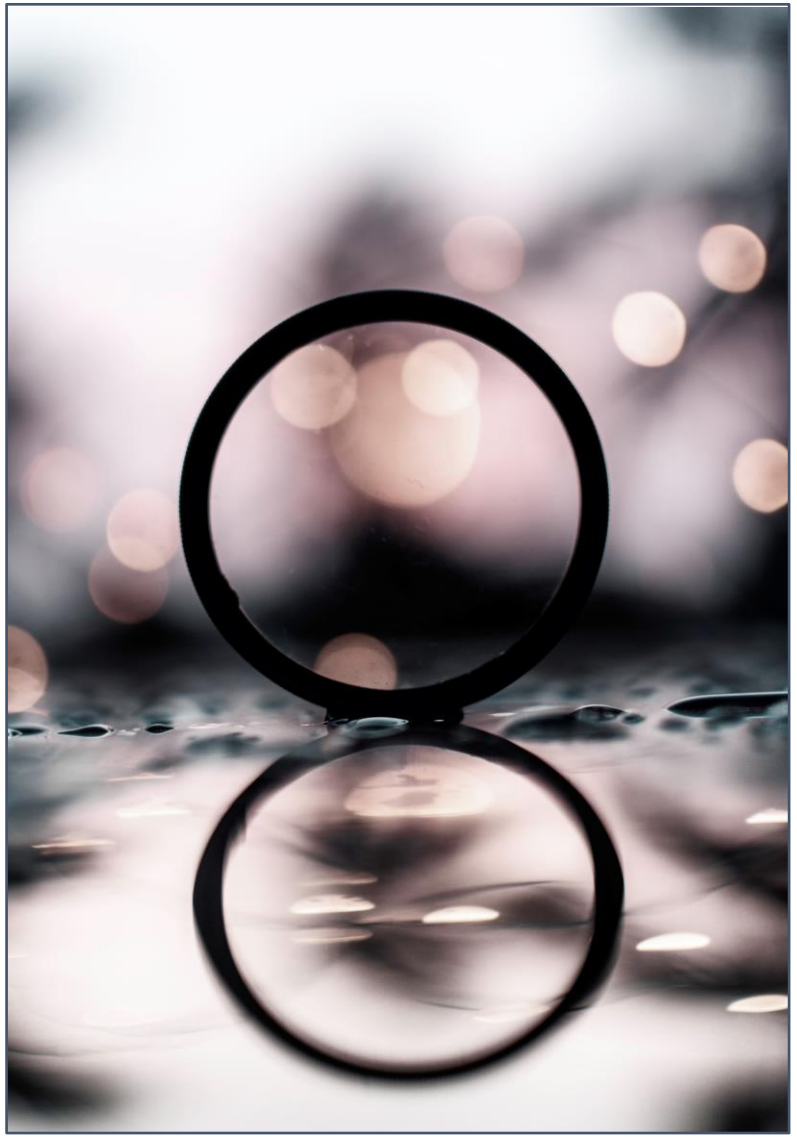
- ¿Cómo determinar las **“semillas”**?
- Fijarla de **modo aleatorio**. No es el mejor método. De hecho, cada vez que se realiza, podrán obtenerse soluciones diferentes.
- **Fijación por parte del investigador**. Una idea, por ejemplo, es hacer un clúster jerárquico previo, y tomar los “centroides” de la solución final como “semillas” de k-medias.
- También existe el método del ***centroide más lejano*****:** se fija el primer centroide al azar, pero luego el 2º centroide coincidirá con el punto de datos más alejado de él. En general, el jº centroide coincidirá con el punto cuya distancia mínima a los centroides precedentes sea mayor. Se pretende que los centroides estén bien separados unos de otros. Una versión mejorada es el método **k-medias++**.

Caracterización de grupos.



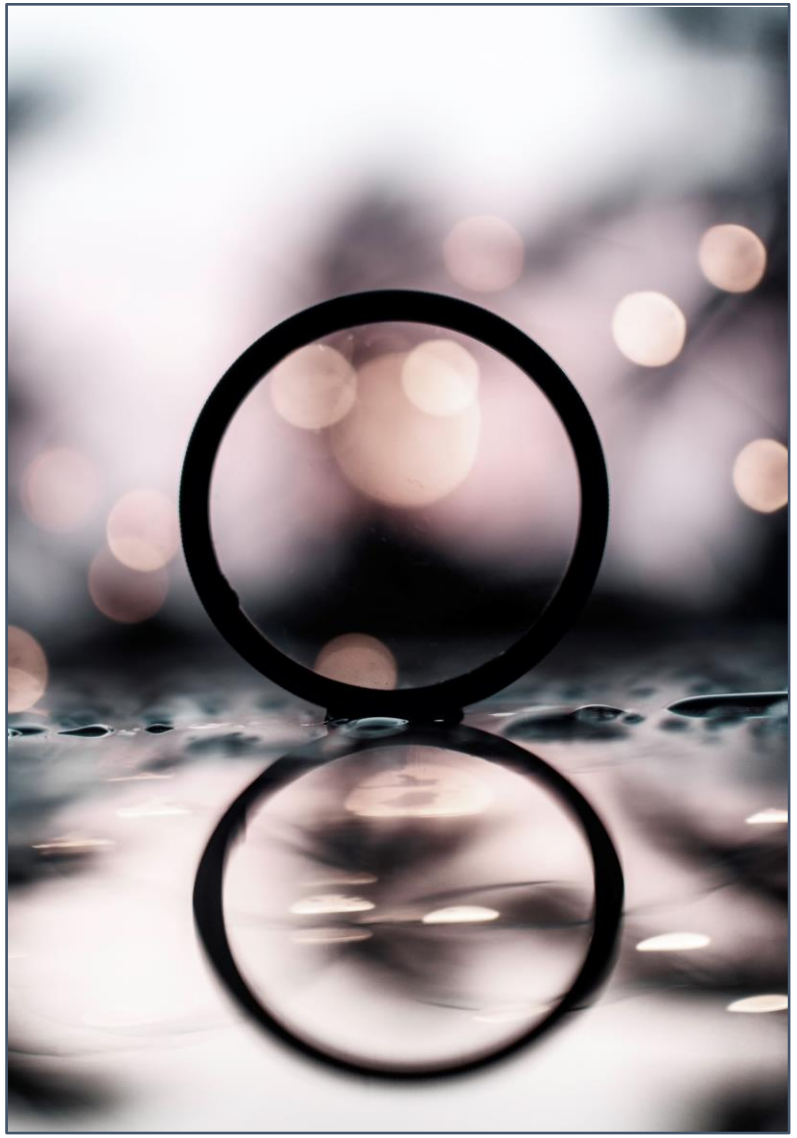
- Una vez establecidos los conglomerados, clústeres o grupos; es importante caracterizarlos para **discernir en qué se distinguen unos de los otros** principalmente.
- Una opción sencilla es establecer las coordenadas de los centroides de cada grupo (en función de las variables clasificadoras; pero sin 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.

Caracterización de grupos.



- Una opción es someter a los grupos, para cada variable, a un **método de comparaciones múltiples de medias**, menos exigente que el ANOVA clásico en cuanto al comportamiento de los datos, como por ejemplo, el **test de comparaciones múltiples de Kruskal-Wallis**.

Caracterización de grupos.



¡Muchas gracias!