Data Science para Economía y Negocios

ANÁLISIS DE CLÚSTER

JAVIER FERNÁNDEZ Y ESTEBAN LÓPEZ

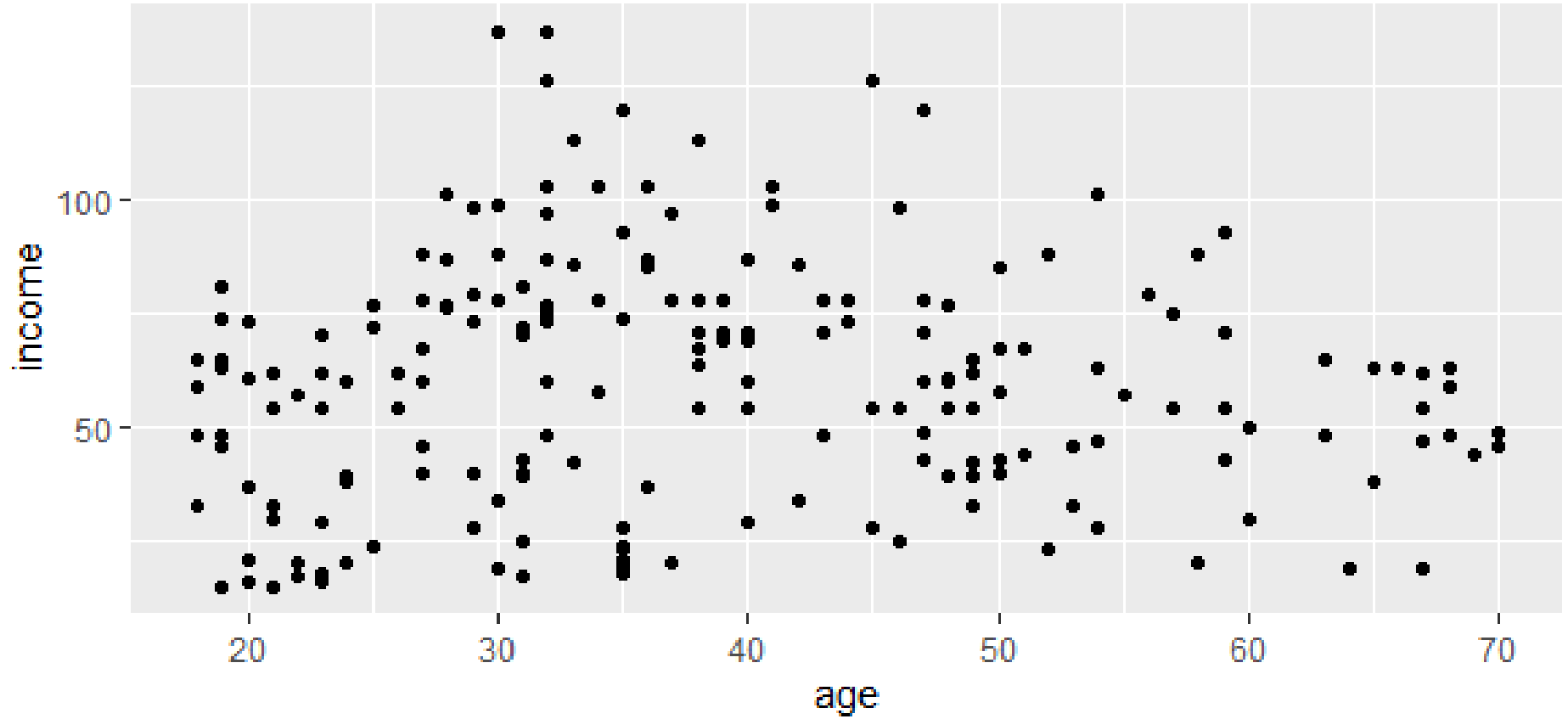
Motivación

- Como hemos visto a lo largo del curso, el análisis de datos es útil en muchas disciplinas (Salud, PP, deportes, etc).
- En particular, en negocios, hemos visto que es una herramienta para agregar valor a las empresas facilitando la toma de decisiones.

Motivación

- Uno de los objetivos de DS es encontrar cierta heterogeneidad en los datos.
- ¿Cómo podemos dividir nuestros datos en subgrupos cuando no existe una variable para clasificarlos?
- Por un momento imagine que trabaja para el departamento de marketing de un empresa…

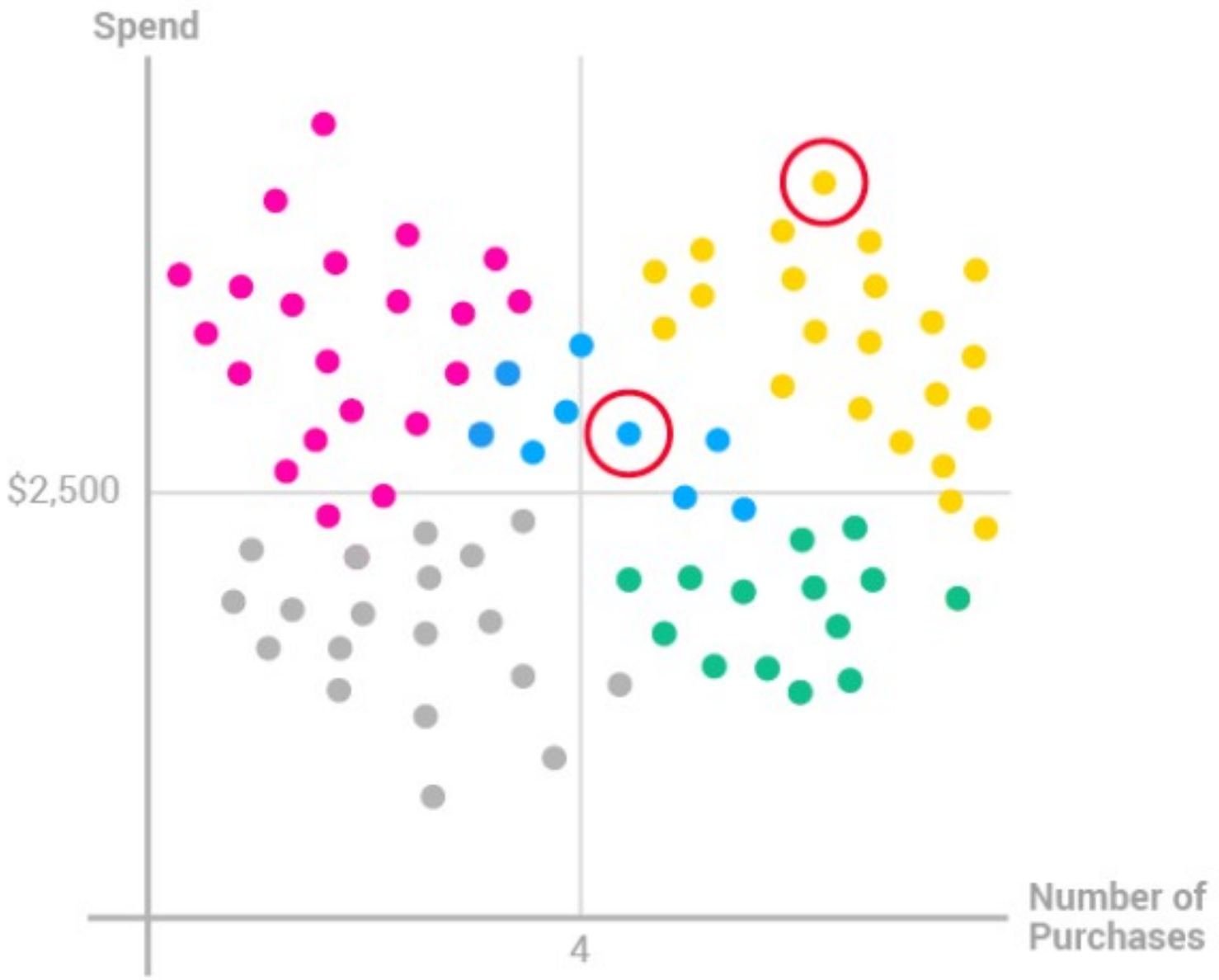
Los datos



¿Cómo capturar la “heterogeneidad” en los datos?

- Existen varias técnicas, las dos más utilizadas en la práctica son análisis de componentes principales(PCA) y análisis de cluster.
  - *Ambos son dos métodos estadísticos data-driven (dejamos a los datos hablar).*
  - *Apuntan a distintos fines:*
    - *PCA busca reducir la dimensionalidad, es decir, reducir el número de variables de importancia en nuestra base de datos.*
    - ***Los métodos de clusterización buscan separar en distintos subgrupos las observaciones de la base de datos.***

Análisis de Cluster



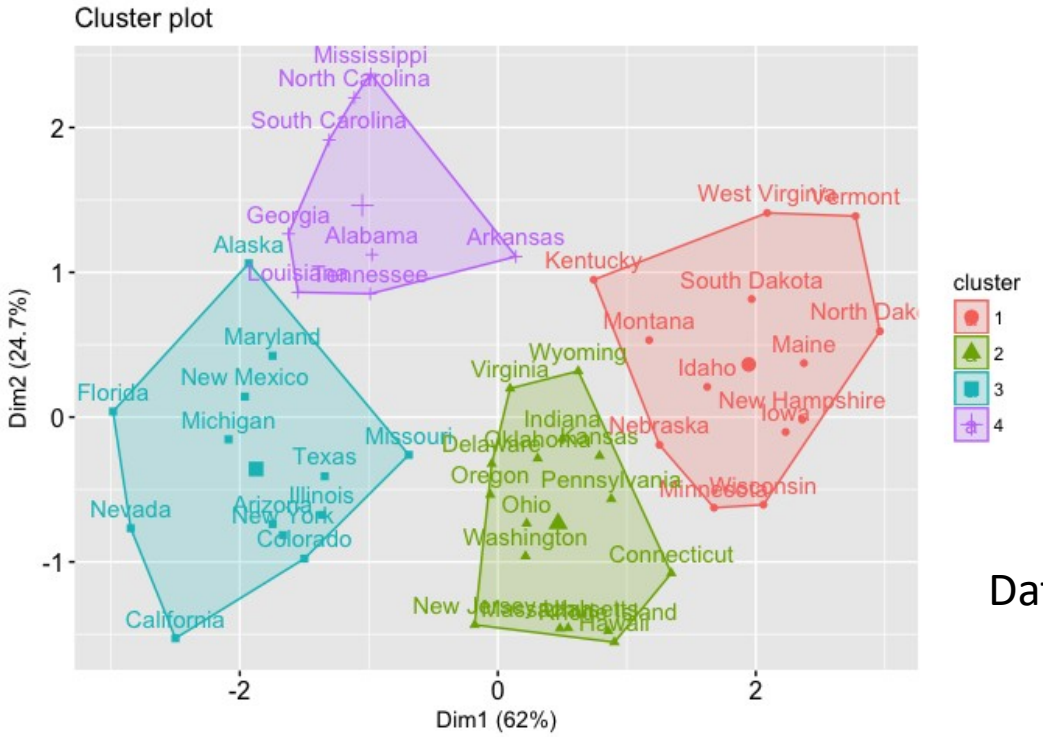
Fuente: optimove.com

Análisis de Cluster

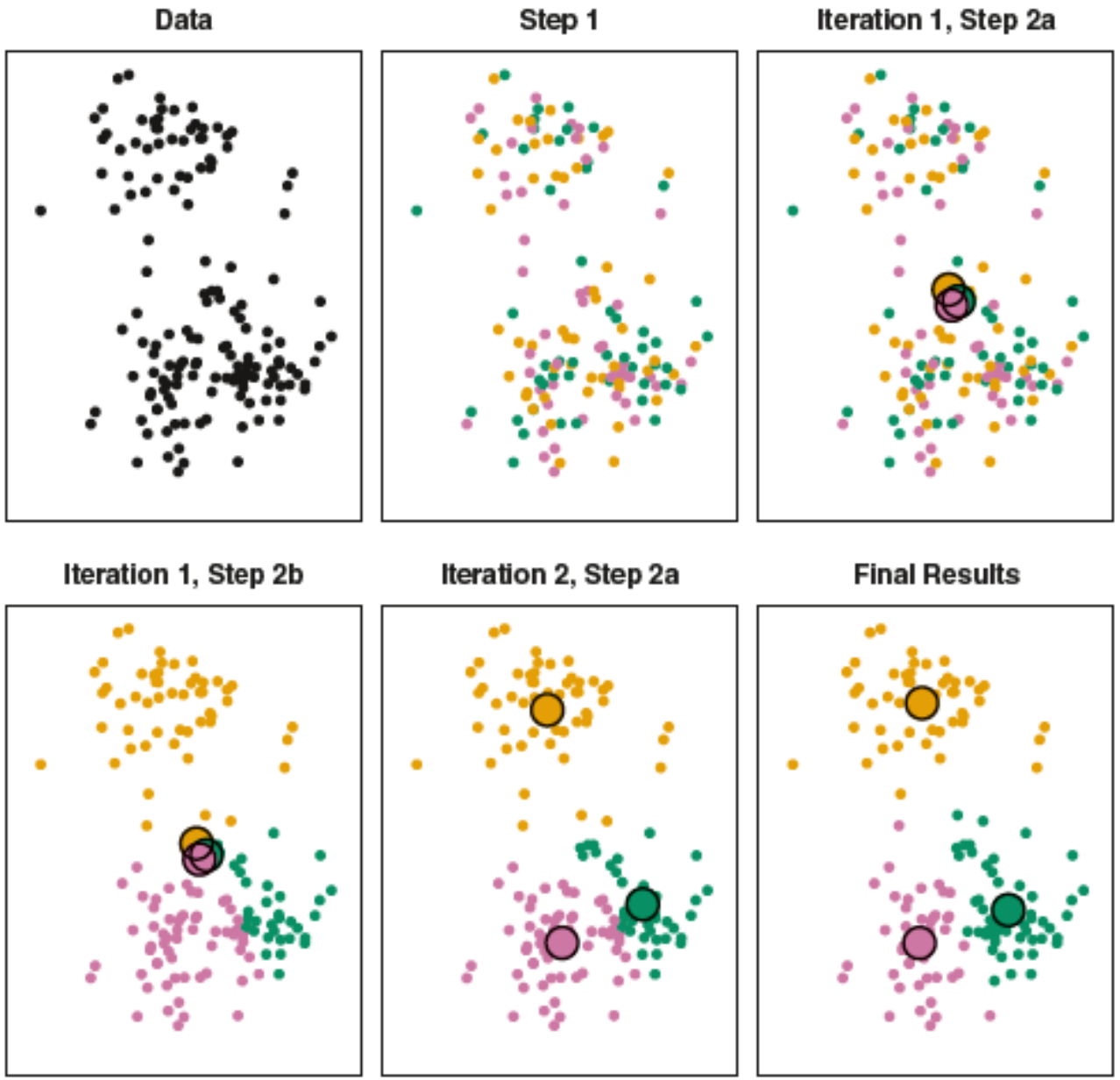
- Los métodos de clusterización tienen como finalidad encontrar cierta heterogeneidad en los datos que nos permitan separar en subgrupos nuestra muestra.
- Existen diferentes tipos (ejemplos):
  - Basados en centroides: **k-means.**
  - Basados en densidad: **DBSCAN**, OPTICS ,etc.
  - Basados en distribución: EM.

K- means

- La idea detrás de este método es que, después de elegir la cantidad de cluster, la asignación a cada cluster de las observaciones sea al que este más cercano **de la media**.
- En términos un poco más estricto el algoritmo iterará hasta que se minimice la distancia de cada observación a cada **centroide** de cada cluster.



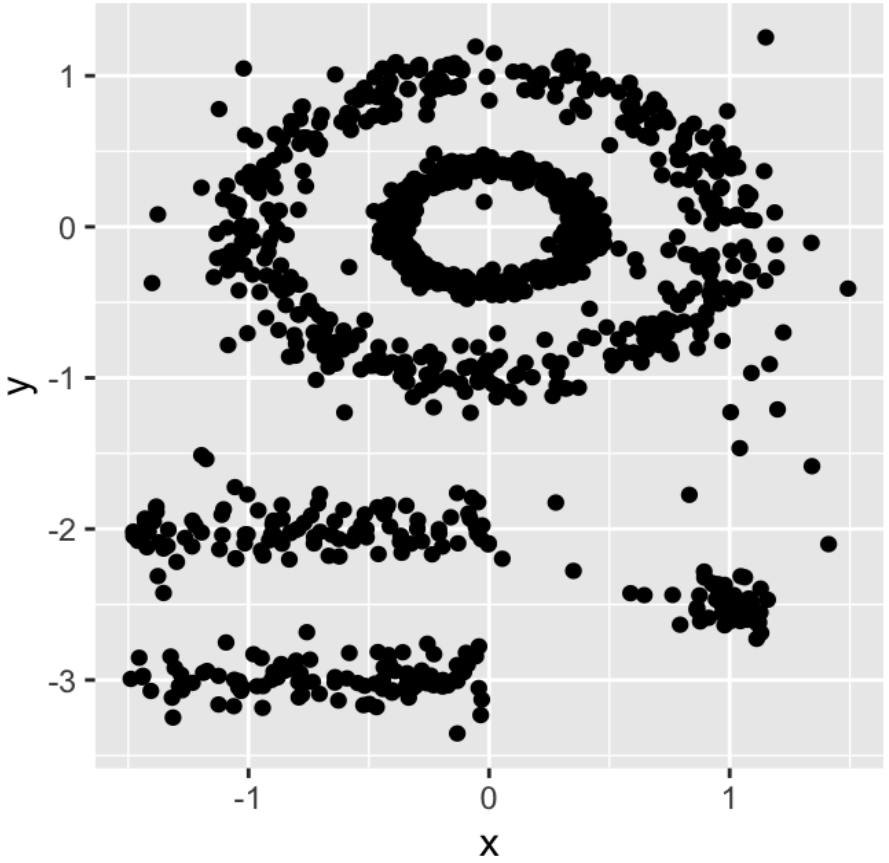
Data: USArrest



Fuente: An Introduction to Statistical Learning, pag. 389.

DBSCAN

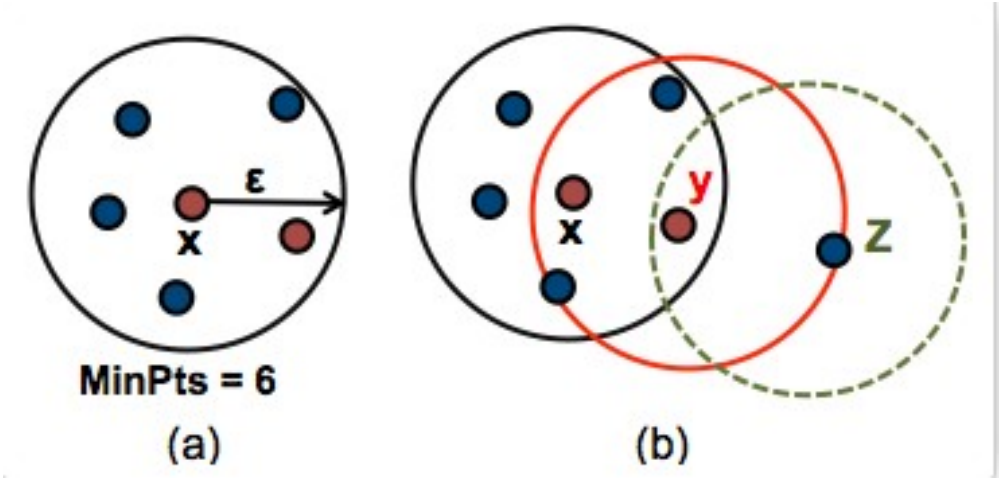
- Cuando los cluster están bien separados, es normal encontrar cluster que gráficamente se ven esféricos o convexos. Pero ¿Cuándo esto no ocurre?



Data: Multishapes

DBSCAN

- Este método de clusterización nos permite encontrar cluster no convexos.
- Es un método basado en densidad, ya que resalta la idea de las vecindades más pobladas.
- En términos simples, el algoritmo crea vecindades dada un radio épsilon y un mínimo de observaciones que deben pertenecer a la vecindad. El cluster se formará por todas las vecindades que estén conectadas.



Fuente: datanovia.com

Lab 11

- Utilizaremos los paquetes “**factoextra**” y **“dbscan”** para realizar análisis de cluster. Las funciones a utilizar son las siguientes:
  - kmeans: permite clusterizar una base de datos mediante el método kmeans. Argumentos:
    - x: base de datos.
    - centers: número de clusters.
    - nstart: número de aleatorizaciones iniciales.
  - dbscan: permite clusterizar una base de datos mediante el método dbscan. Argumentos:
    - x: base de datos.
    - eps: radio de definición de la vecindad.
    - minPts: número de puntos mínimos para constituir una vecindad. (Por defecto 3).
  - kNNdistplot: permite visualizar la distancia promedio a los k vecinos más cercano de cada observación de la base de datos.
    - x: base de datos.
    - k: número de vecinos a considerar.
  - fviz_cluster: permite visualizar gráficamente los cluster.
    - object: objeto tipo cluster que se quiere visualizar
    - data: base de datos.