www.nataliejulian.com
Natalie Julian

Ejercicios: Clustering con K-means

Contexto

Es bien sabido que los costos de vida varían en todo el mundo. Una analista económica está interesada en analizar los costos de vida alrededor de todo el mundo y en particular, le interesa establecer cuáles son las ciudades que, en términos de costos de vida, son símiles entre ellas y también cuáles no son símiles. La base de datos `cost of living` contiene la siguiente información:

- `City` Nombre de la ciudad y el país
- `Cost of Living Index` Indicador relativo de los precios de los bienes de consumo, incluidos comestibles, restaurantes, transporte y servicios públicos en la ciudad, no incluye los gastos de alojamiento, como alquiler o hipoteca
- `Rent Index` Indicador relativo de los precios de arriendos o alquileres en la ciudad
- `Cost of Living Plus Rent Index` Indicador relativo de los precios de bienes de consumo y rentas en la ciudad
- `Groceries Index` Indicador relativo de los precios de comestibles en la ciudad
- `Restaurant Price Index` Indicador relativo de los precios de comidas y bebidas en restaurantes y bares en la ciudad
- `Local Purchasing Power Index` Indicador relativo del poder adquisitivo de los residentes en la ciudad con un sueldo promedio en la ciudad

(*) Todos los índices usaron como referencia Nueva York, es decir, para la ciudad de Nueva York, cada índice es 100. Si otra ciudad tiene, por ejemplo, un índice de 120, significa que en promedio los precios son un 20% más caro que en la ciudad de Nueva York. Si una ciudad tiene un índice de 70, eso significa que, en promedio, los precios en esa ciudad son 30% menos costosas que en Nueva York.

Desarrollo

i) Realice clustering con k-means. Justifique cada uno de los pasos que realiza para determinar la clusterización final. Utilice como semilla 2019. Recuerde:

- Analizar previamente las variables a utilizar en k-means (naturaleza, escalas de medición, variabilidad, etcétera)
- Considerar tratamientos previos a los datos (análisis de outliers (utilice el criterio de detección que prefiera), estandarización, etcétera)
- Probar distintos valores de k y proponer en base a criterios explícitos, posibles valores de k a utilizar y cuál elegiría usted
- Realizar la clusterización y comentar detalles importantes a la hora de analizar la segmentación obtenida

ii) Añada la clusterización encontrada a la base de datos. ¿En qué clúster se encuentra la ciudad Santiago, Chile? ¿qué otras ciudades presentan un comportamiento (en términos de costos de vida) similar a Santiago de Chile? (nombre 3)