Expo: Estadística en el Entorno

"Explorando la Criminalidad: Silhouette y k-means en EE. UU."

Participantes:

Diego Felipe Hernández Justo, Miguel Angel Anguiano Pérez, María Luisa Córdoba Tlaxcalteco.

Introducción

En el mundo de la estadística multivariada, destacamos la importancia del método no supervisado k-means, centrándonos específicamente en su aplicación a la base de datos USArrest.
Este algoritmo de agrupamiento se convierte en una herramienta esencial para simplificar la complejidad de grandes conjuntos de datos, dividiendo la información en clústeres definidos. A través de esta exploración, buscamos comprender las variaciones en las tasas de arresto y las dinámicas regionales, ofreciendo una visión más clara y detallada de la complejidad de la criminalidad en diferentes partes del país.

Base

La base de datos USArrest presenta estadísticas detalladas sobre arrestos en cada uno de los 50 estados de EE.UU. Las cifras se presentan por cada 100,000 residentes y abarcan distintas categorías delictivas, incluyendo asesinato, asalto y violación. Cada estado se identifica por su nombre y se proporciona información específica sobre las detenciones por asesinato (Murder), asalto (Assault), violación (Rape), y el porcentaje de población urbana (UrbanPoP). Estos datos ofrecen una visión integral de las dinámicas criminales a nivel estatal, permitiendo un análisis detallado de las tendencias delictivas en distintas regiones del país.

Metodología

1. Selección de la Base de Datos:
Utilizamos la base de datos "USArrest" de RStudio, que contiene información crucial sobre arrestos en los 50 estados de EE. UU.

2. Determinación del Número Óptimo de Clústeres (k):
Aplicamos el método de Silhouette para identificar el valor óptimo de k (número de clústeres) necesario para un agrupamiento significativo de los datos de arrestos.

3. Interpretación de los Clústeres:
Analizamos detalladamente los clústeres generados por el método k-means, explorando las características específicas de cada grupo para identificar patrones y distinciones en las dinámicas de arrestos en los diferentes estados.

Discusión

La identificación del codo en el Gráfico 1 sugiere que dos clusters son óptimos, respaldando la eficiencia del método de Silhouette.
En el Gráfico 2, los estados en el Clúster 1 muestran perfiles asociados a mayor seguridad. La coincidencia entre la recomendación de dos clusters y la visualización valida la robustez del análisis, proporcionando una comprensión clara y diferenciada de las dinámicas criminales en los estados.
En resumen, la combinación de Silhouette y k-means permite una interpretación efectiva de patrones significativos en la seguridad pública con un enfoque de dos clusters.

Resultados

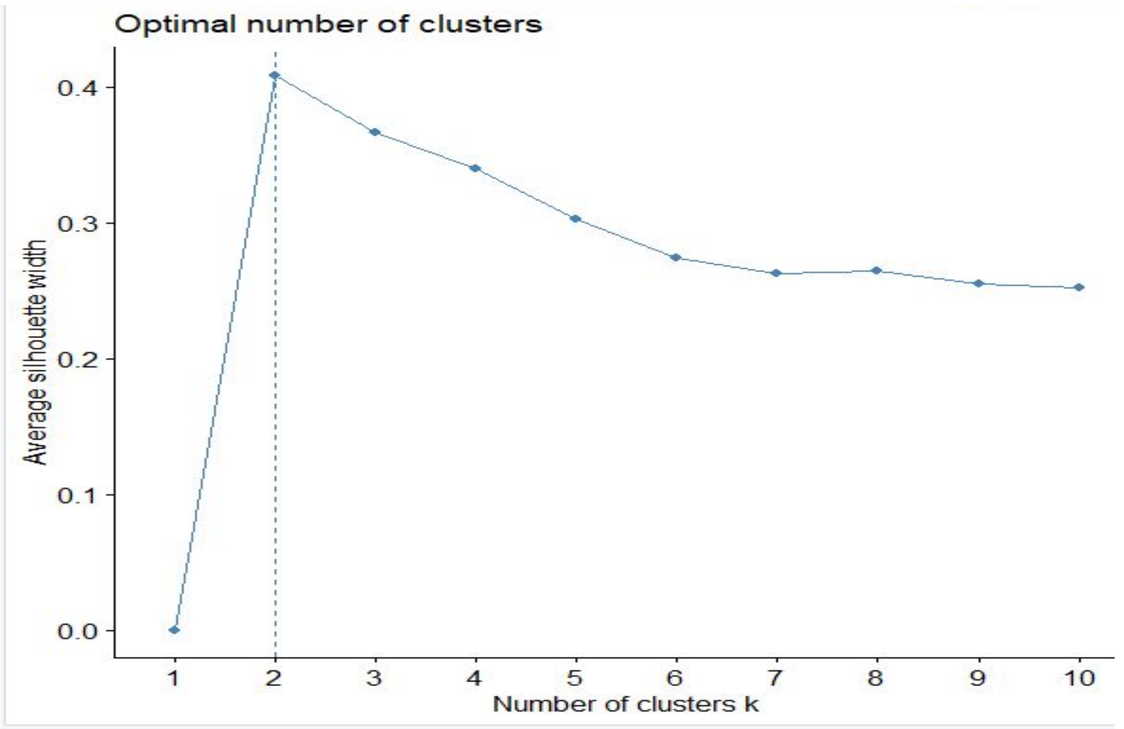

Gráfico 1. Representación del número de Clúster.

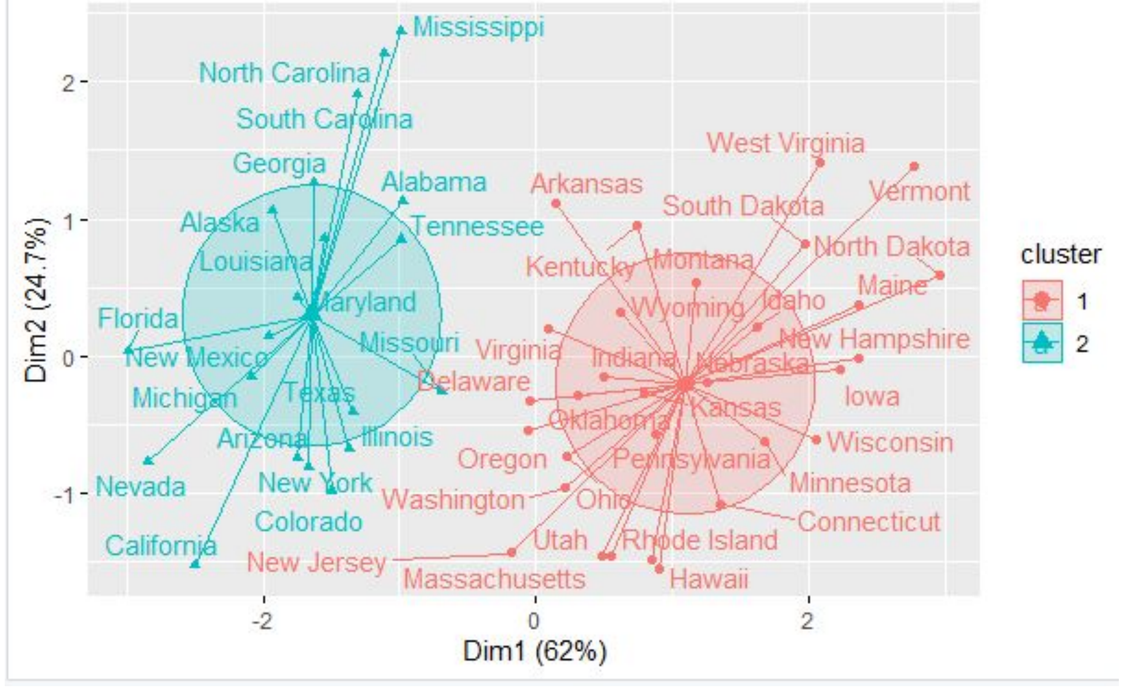

Gráfico 2. Representación del número de Cluster Plot.

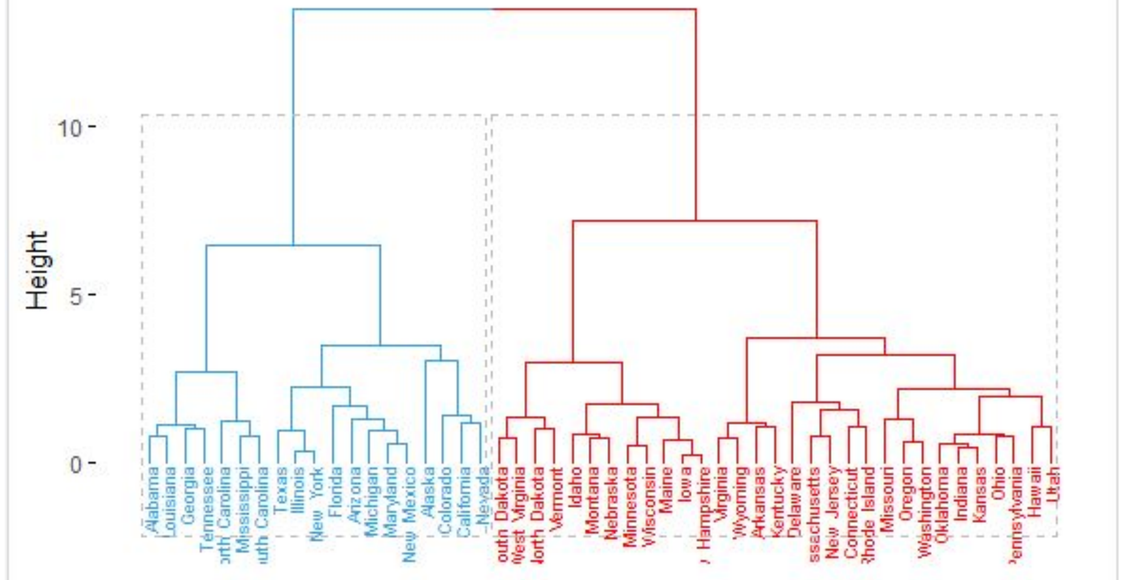

Gráfico 3. Representación del dendograma de Cluster.

Referencias

1. R Dataset / Package Datasets / USARRests | R Datasets. (s. f.). Recuperado de https://r-data.pmagunia.com/dataset/r-dataset-package-datasets-usarrests
2. El algoritmo k-means aplicado a clasificación y procesamiento de imágenes. (s. f.). Universidad de Oviedo. Recuperado de https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html
3. Marrero, L., Carrizo, D., García-Santander, L., & Ulloa-Vásquez, F. (2021). Uso de algoritmo K-means para clasificar perfiles de clientes con datos de medidores inteligentes de consumo eléctrico: Un caso de estudio. Ingeniare. Revista chilena de ingeniería, 29(4), 778-787.