INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

Carrera: Ingeniería en Inteligencia Artificial

Unidad de aprendizaje: Fundamentos de Inteligencia Artificial

Práctica 10: Aplicando K-means

Alumnos:

Caballero Chávez Yael Jesús
Dominguez Rendon Melissa

Grupo:
4BM1

Profesor: Hernández Cruz Macario

RESUMEN

En esta práctica, se aborda la implementación de un algoritmo de clustering K-means aplicado a un conjunto de datos globales de criminalidad, específicamente utilizando un archivo denominado global_index.csv. Este archivo contiene información sobre diferentes países, con columnas que representan diversos aspectos de criminalidad: "Criminality", "Criminal markets" y "Criminal actors". El objetivo principal es agrupar los países en distintos clusters según sus características de criminalidad, utilizando el algoritmo de K-means, y posteriormente visualizar estos clusters en un mapa geoespacial.

INTRODUCCIÓN

El clustering K-means es una técnica de aprendizaje no supervisado que se utiliza para particionar un conjunto de datos en K grupos o clusters, donde cada grupo contiene puntos de datos similares entre sí. En esta práctica, se busca encontrar el número óptimo de clusters K que mejor describa la estructura del conjunto de datos de criminalidad. Para ello, se emplean métricas de evaluación como el coeficiente de silueta, el índice de Davies-Bouldin y el índice de Calinski-Harabasz.

Además de la implementación del algoritmo K-means y la evaluación de su rendimiento, se incluye la visualización de los resultados en un mapa geoespacial utilizando la biblioteca GeoPandas. Este mapa permitirá identificar visualmente cómo se distribuyen los diferentes niveles de criminalidad entre los países agrupados, proporcionando una herramienta visual poderosa para el análisis geoespacial de los datos.

Esta práctica combina el aprendizaje automático con la visualización de datos geoespaciales, ofreciendo una comprensión más profunda de los patrones de criminalidad a nivel global y las posibles relaciones entre diferentes aspectos de la misma. La implementación de estas técnicas se realiza en el lenguaje de programación Python, utilizando bibliotecas como pandas, scikit-learn, GeoPandas y matplotlib, entre otras.

DESARROLLO

Lo primero a hacer fue poder encontrar un K óptimo para poder hacer la agrupacion de la mejor forma. Para ello utilizamos un programa de Python que nos arroja la siguiente grafica donde nos indica el k = 3.

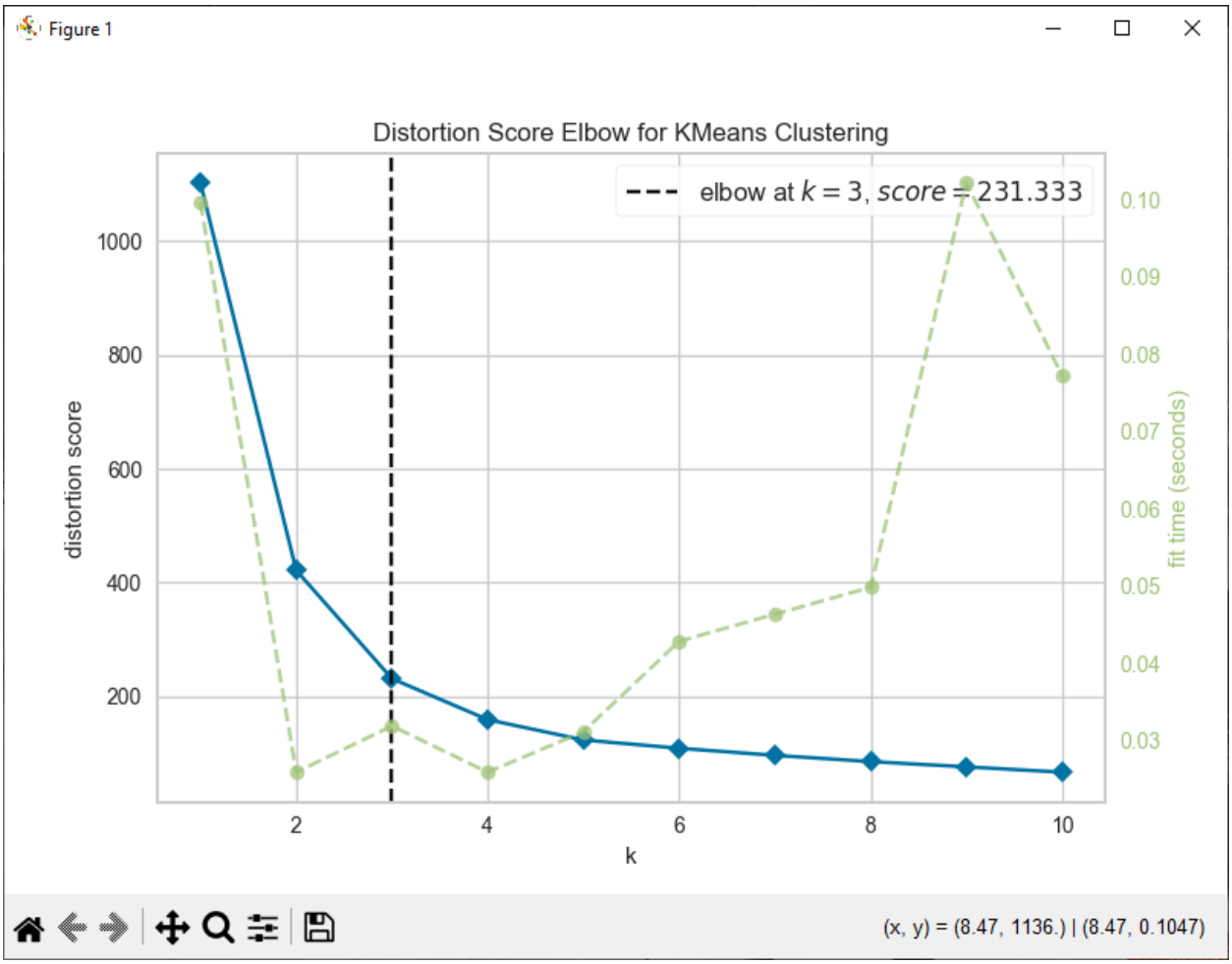


Figure 1
Distortion Score Elbow for KMeans Clustering
elbow at $k = 3$, $score = 231.333$
distortion score
fit time (seconds)
k
(x, y) = (8.47, 1136.) | (8.47, 0.1047)

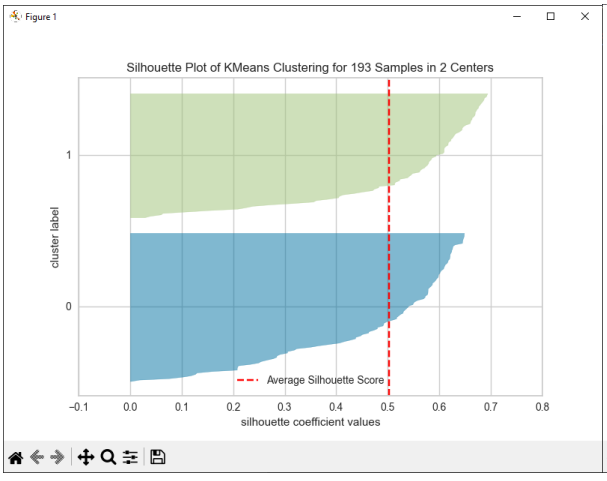


Figure 1
Silhouette Plot of KMeans Clustering for 193 Samples in 2 Centers
cluster label
Average Silhouette Score
silhouette coefficient values

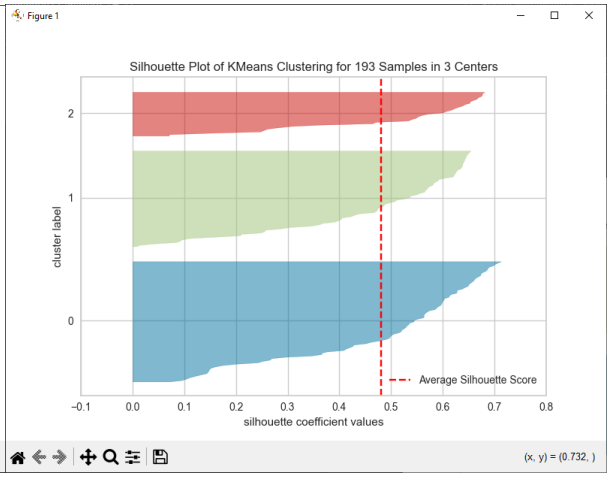


Figure 1
Silhouette Plot of KMeans Clustering for 193 Samples in 3 Centers
cluster label
Average Silhouette Score
silhouette coefficient values
(x, y) = (0.732,)

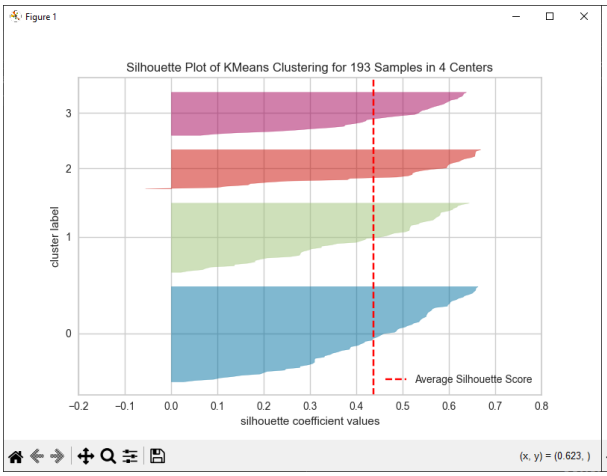


Figure 1
Silhouette Plot of KMeans Clustering for 193 Samples in 4 Centers
cluster label
silhouette coefficient values
Average Silhouette Score
(x, y) = (0.623,)

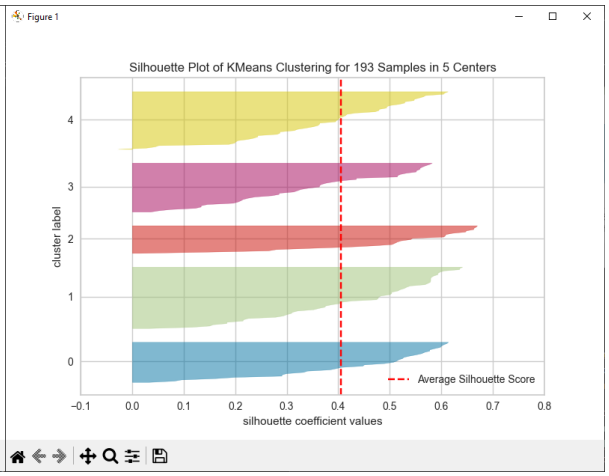


Figure 1
Silhouette Plot of KMeans Clustering for 193 Samples in 5 Centers
cluster label
silhouette coefficient values
Average Silhouette Score

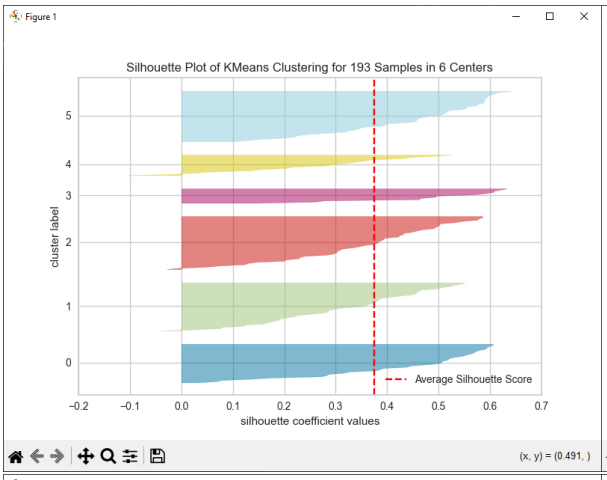


Figure 1
Silhouette Plot of KMeans Clustering for 193 Samples in 6 Centers
cluster label
silhouette coefficient values
Average Silhouette Score
(x, y) = (0.491,)

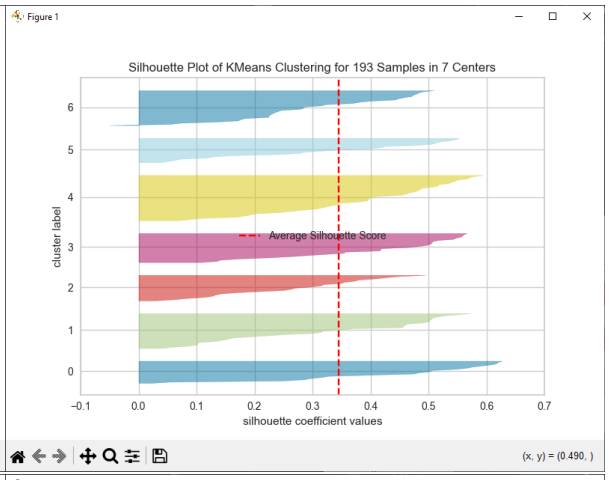


Figure 1
Silhouette Plot of KMeans Clustering for 193 Samples in 7 Centers
cluster label
silhouette coefficient values
Average Silhouette Score
(x, y) = (0.490,)

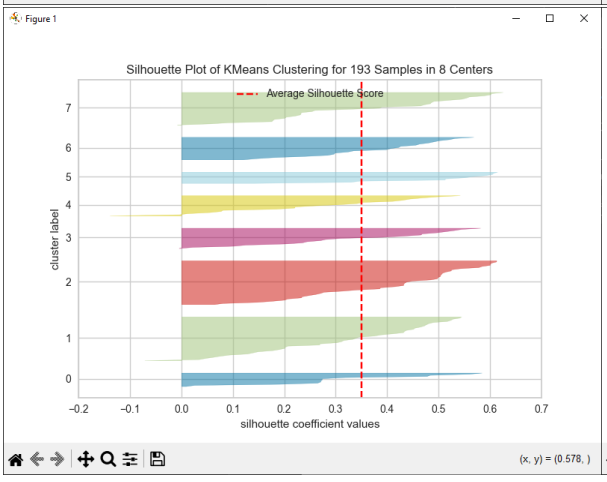


Figure 1
Silhouette Plot of KMeans Clustering for 193 Samples in 8 Centers
Average Silhouette Score
cluster label
silhouette coefficient values
(x, y) = (0.578,)

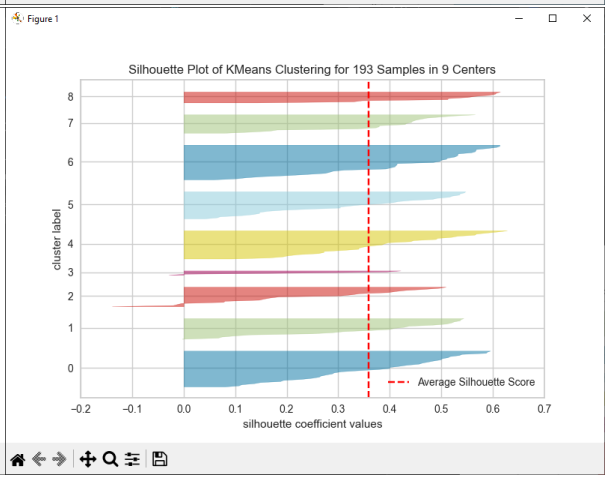
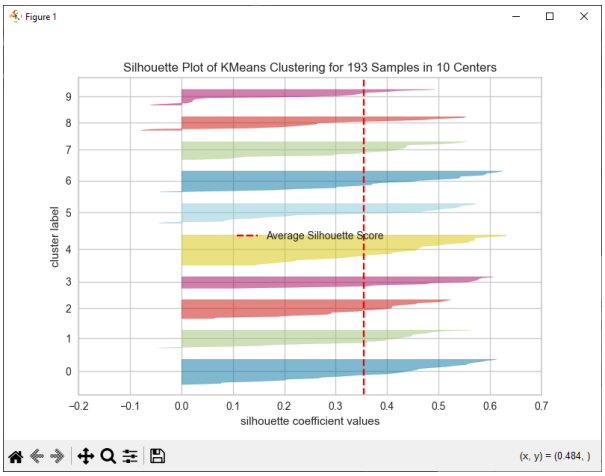
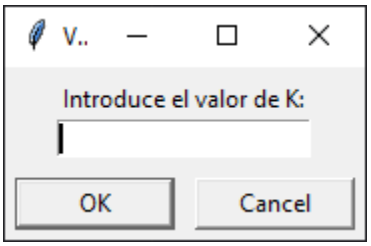


Figure 1
Silhouette Plot of KMeans Clustering for 193 Samples in 9 Centers
cluster label
silhouette coefficient values
Average Silhouette Score

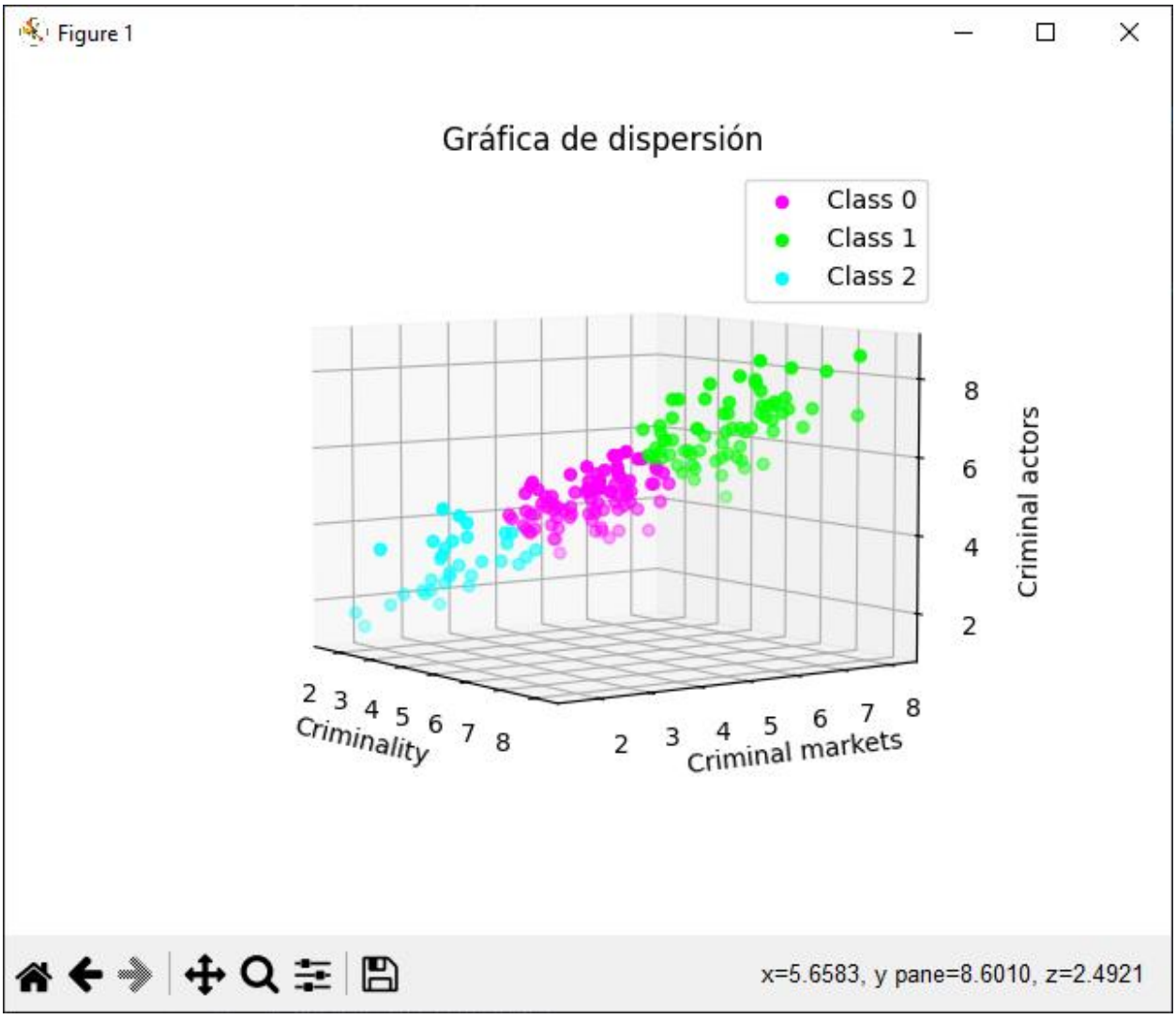


Una vez obtenido el k optimo igual a 3, podemos agrupar los países.

Cuando ejecutamos el programa nos pide ingresar el k que ya conocemos.



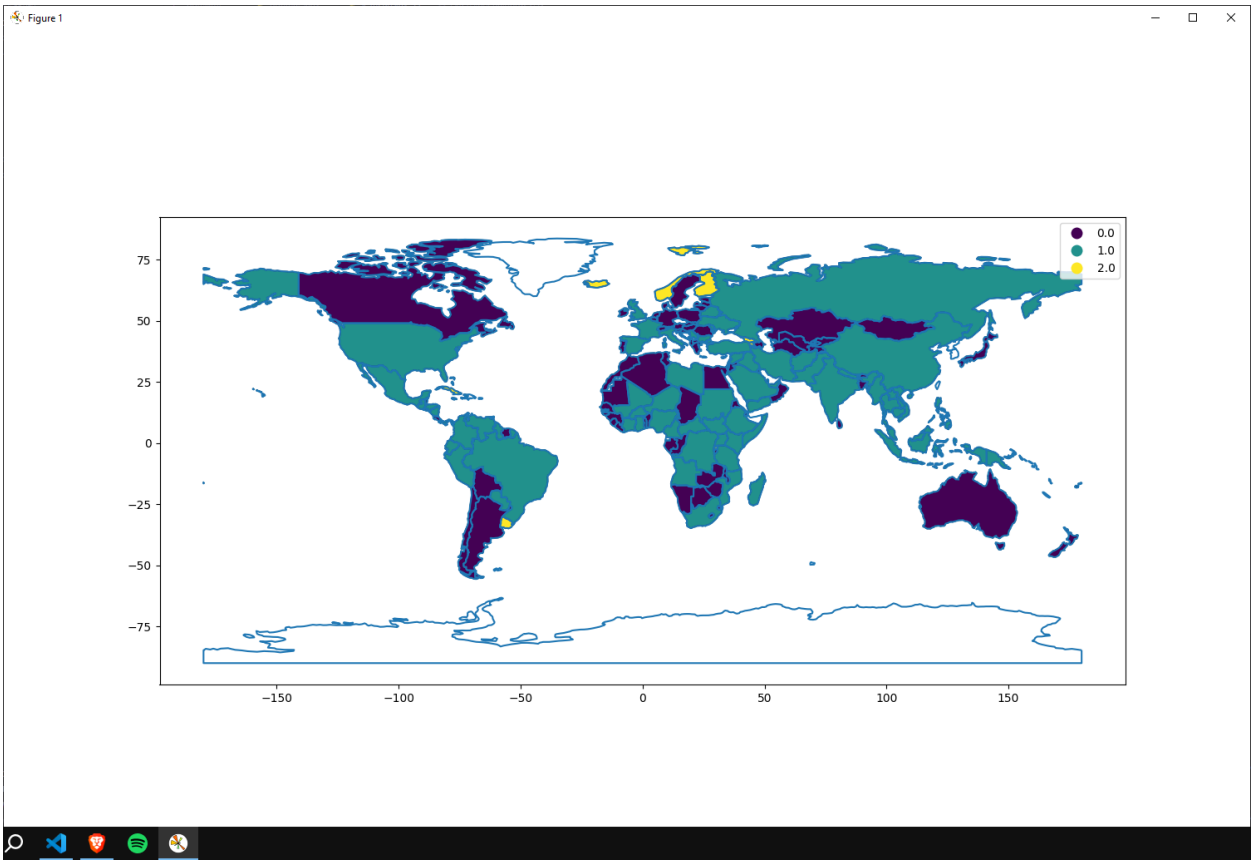
Al ingresarlo nos despliega inmediatamente la gráfica con todos los países agrupados.



DISCUSIONES DE LOS RESULTADOS

Country	Criminality	Criminal ma	Criminal ac	Class
Timor-Leste	4.08	3.67	4.5	0
Djibouti	4.65	4.3	5	0
Algeria	4.88	5.17	4.6	0
Fiji	4.15	4.3	4	0
Botswana	4.35	4.4	4.3	0
Kuwait	5.2	5.7	4.7	0
Mauritius	4.37	4.13	4.6	0
Jordan	4.93	4.87	5	0
Maldives	4.27	3.53	5	0
Sierra Leon	4.95	4.6	5.3	0
Bangladesh	5.12	5.03	5.2	0
Costa Rica	5.53	5.37	5.7	0
Senegal	5.52	5.53	5.5	0
Lesotho	3.92	3.43	4.4	0
Poland	4.48	4.97	4	0
Uzbekistan	4.95	4.6	5.3	0
Croatia	5.15	4.9	5.4	0
Solomon Isl	4.4	3.7	5.1	0
Egypt	5.05	5.1	5	0
Kyrgyzstan	5.32	4.63	6	0
Argentina	5	4.5	5.5	0
Azerbaijan	4.8	4.1	5.5	0
Romania	4.58	5.27	3.9	0
Cyprus	4.43	3.97	4.9	0
Togo	5.23	4.77	5.7	0
Malta	5	4.3	5.7	0
Greece	5.35	4.7	6	0
Gambia	4.53	4.67	4.4	0
Oman	4.4	4.9	3.9	0
Mongolia	4.12	3.83	4.4	0
Hungary	4.62	4.73	4.5	0
Seychelles	3.9	3.5	4.3	0
Bolivia	4.95	5	4.9	0
Dominican	5.02	5.13	4.9	0
Bahrain	4.95	5.4	4.5	0
North Mace	5.03	4.87	5.2	0
Bhutan	3.9	3.9	3.9	0
Trinidad an	5.2	4.8	5.6	0
Morocco	4.8	5.1	4.5	0
Malawi	4.48	4.77	4.2	0
Namibia	4.3	4.1	4.5	0
Kazakhstan	4.47	4.33	4.6	0
Zambia	4.73	4.47	5	0
Qatar	5.45	5.7	5.2	0
Tunisia	4.45	5	3.9	0
Slovakia	4.72	4.73	4.7	0
Benin	5.32	5.43	5.2	0
Albania	5.17	4.83	5.5	0
Israel	4.85	5	4.7	0
Slovenia	4.37	4.03	4.7	0
Ireland	5.08	5.17	5	0
Guinea-Bis	5.1	4.6	5.6	0
Tajikistan	5.45	4.8	6.1	0
Japan	4.28	3.87	4.7	0
Chad	5.5	5.1	5.9	0
Comoros	3.92	3.73	4.1	0
Australia	4	4.3	3.7	0
Netherland	4.97	5.23	4.7	0
Eritrea	3.97	3.93	4	0
Lithuania	3.9	3.8	4	0
Equatorial (	4.38	3.57	5.2	0
Burundi	4.87	4.63	5.1	0
Sweden	4.7	4.6	4.8	0
Germany	5.33	5.47	5.2	0
Austria	4.13	4.17	4.1	0
Korea, DPR	4.82	5.73	3.9	0
Latvia	3.9	4.5	3.3	0
Estonia	4.25	4.2	4.3	0
New Zealar	4.08	3.77	4.4	0
Korea, Rep,	4.43	3.57	5.3	0
Denmark	4.02	4.33	3.7	0
Turkmenist	4.4	4.4	4.4	0
Zimbabwe	5.47	5.03	5.9	0
Suriname	4.77	4.53	5	0
Mauritania	4.38	4.27	4.5	0
Guinea	4.58	4.77	4.4	0
Cabo Verde	4.28	3.97	4.6	0
Portugal	4.88	4.67	5.1	0
Congo	4.78	4.47	5.1	0
Gabon	4.85	4.6	5.1	0
eSwatini	4.38	3.87	4.9	0
Belize	4.87	4.43	5.3	0
Liberia	5.5	5.4	5.6	0
Switzerlanc	4.87	4.63	5.1	0
Belgium	4.43	5.17	3.7	0
Czech Repu	4.68	4.87	4.5	0
Canada	3.88	3.87	3.9	0
Chile	5.18	5.07	5.3	0
Sri Lanka	4.92	4.83	5	0
Malaysia	6.23	6.67	5.8	1
United King	5.75	5.5	6	1
Italy	6.22	5.73	6.7	1
France	5.82	5.93	5.7	1
Kenya	7.02	6.93	7.1	1
Bulgaria	5.65	5.4	5.9	1
Nigeria	7.28	7.37	7.2	1
China	6.37	6.53	6.2	1
Jamaica	5.8	4.9	6.7	1
India	5.75	6.7	4.8	1
South Africa	7.18	6.87	7.5	1
Colombia	7.75	7.3	8.2	1

United Stat	5.67	5.83	5.5	1
Ghana	5.8	6	5.6	1
Spain	5.9	5.7	6.1	1
Afghanistan	7.1	7	7.2	1
Ecuador	7.07	6.73	7.4	1
United Arab	6.37	7.03	5.7	1
Tanzania	6.2	6.4	6	1
Sudan	6.37	5.23	7.5	1
Guatemala	6.6	6.1	7.1	1
Honduras	7.05	6	8.1	1
Guyana	5.97	5.13	6.8	1
Iran	7.03	7.37	6.7	1
Saudi Arabi	6.23	6.57	5.9	1
Cameroon	6.27	6.23	6.3	1
Nepal	6.57	6.03	7.1	1
Pakistan	6.03	6.27	5.8	1
Iraq	7.13	6.27	8	1
Côte d'Ivoir	6.02	5.93	6.1	1
Haiti	5.93	5.77	6.1	1
Moldova	5.6	5.2	6	1
Bosnia and	5.85	5.3	6.4	1
Uganda	6.55	6.4	6.7	1
Russia	6.87	6.83	6.9	1
Cambodia	6.85	6.7	7	1
Niger	5.7	5.7	5.7	1
Burkina Fas	5.92	5.83	6	1
Belarus	5.87	5.33	6.4	1
Mozambiqu	6.2	5.9	6.5	1
Laos	6.12	6.33	5.9	1
Lebanon	7.1	6.3	7.9	1
Paraguay	7.52	6.73	8.3	1
Turkey	7.03	6.77	7.3	1
El Salvador	5.92	5.43	6.4	1
Mexico	7.57	8.13	7	1
Philippines	6.63	6.57	6.7	1
Mali	5.93	6.47	5.4	1
Libya	6.93	6.57	7.3	1
Myanmar	8.15	7.7	8.6	1
Serbia	6.22	5.73	6.7	1
Yemen	6.57	5.63	7.5	1
Brazil	6.77	6.93	6.6	1
Central Afri	6.75	5.6	7.9	1
Somalia	6.13	5.27	7	1
Thailand	6.18	6.77	5.6	1
Vietnam	6.55	6.5	6.6	1
Madagasca	5.58	5.27	5.9	1
Ethiopia	5.68	6.07	5.3	1
Venezuela	6.72	6.03	7.4	1
Panama	6.98	6.67	7.3	1
Indonesia	6.85	6.6	7.1	1
Dem. Rep. (	7.35	6.2	8.5	1
Peru	6.4	6.2	6.6	1
Montenegro	5.9	5.2	6.6	1
Ukraine	6.48	6.27	6.7	1
Angola	5.58	5.17	6	1
Syria	7.07	6.43	7.7	1
S. Sudan	6.32	5.13	7.5	1
Nicaragua	5.72	5.23	6.2	1
Papua New	5.72	5.33	6.1	1
Singapore	3.47	3.93	3	2
Norway	3.75	4.1	3.4	2
Andorra	3.22	2.73	3.7	2
Uruguay	3.22	3.33	3.1	2
Luxembourg	2.85	2.9	2.8	2
Iceland	3.37	2.93	3.8	2
Samoa	2.43	2.97	1.9	2
Tuvalu	1.62	1.93	1.3	2
Nauru	2.05	2.2	1.9	2
Grenada	2.93	2.67	3.2	2
Vanuatu	2.43	2.67	2.2	2
St, Kitts and	3.52	2.83	4.2	2
Dominica	2.63	2.67	2.6	2
Sao Tome a	1.7	1.7	1.7	2
Liechtenste	2.27	2.33	2.2	2
San Marino	3.48	2.37	4.6	2
Georgia	3.6	3.6	3.6	2
Tonga	3.7	3.5	3.9	2
Palau	2.7	2.9	2.5	2
Cuba	3.37	3.63	3.1	2
St, Vincent	3.08	2.67	3.5	2
Brunei	2.85	3.3	2.4	2
Antigua and	2.98	2.67	3.3	2
Kiribati	2.45	2.6	2.3	2
Bahamas	3.75	3.6	3.9	2
Rwanda	3.6	4	3.2	2
St, Lucia	3.53	2.67	4.4	2
Monaco	2.58	1.67	3.5	2
Armenia	2.82	2.93	2.7	2
Marshall Isl	2.52	2.73	2.3	2
Micronesia	3	3	3	2
Barbados	3.07	2.43	3.7	2
Finland	2.98	3.27	2.7	2



Los países del grupo 3 son mucho menos, pero muchos no están en el mapa o no se encuentran bien.

Y algunos otros países como Groenlandia no s encuentran en la lista de datos de criminalidad.

CONCLUSIONES

En esta práctica, se implementó el algoritmo de clustering K-means en un conjunto de datos de criminalidad global, lo que permitió agrupar distintos países según sus características de criminalidad. A través de esta experiencia, se obtuvieron varias conclusiones significativas:

Eficacia del Algoritmo K-means: El algoritmo K-means demostró ser efectivo para identificar y agrupar países con patrones similares de criminalidad. Las métricas de evaluación, como el coeficiente de silueta, el índice de Davies-Bouldin y el índice de Calinski-Harabasz, fueron útiles para determinar la calidad de los clusters formados y seleccionar el número óptimo de clusters K.

Importancia de la Selección del Valor de K: La selección adecuada del valor de K es crucial para el éxito del clustering. El uso de métodos como el codo y la silueta ayudó a identificar el número óptimo de clusters, mejorando así la interpretabilidad y la coherencia de los resultados.

Visualización Geoespacial: La integración de los resultados del clustering con la visualización geoespacial proporcionó una perspectiva clara y comprensible de la distribución global de la criminalidad. La

visualización en el mapa geoespacial destacó las diferencias y similitudes entre los países, facilitando la identificación de patrones y tendencias en los datos de criminalidad.

Desafíos de Coincidencia de Nombres: Uno de los desafíos encontrados fue la coincidencia de nombres de los países entre el conjunto de datos de criminalidad y el conjunto de datos geoespaciales. La limpieza y normalización de los nombres fueron pasos cruciales para asegurar una integración precisa y sin errores de los datos.

Relleno de Valores Faltantes: La asignación de una clase específica para los valores faltantes permitió asegurar que todos los países fueran representados en el mapa, evitando áreas blancas o sin información en la visualización.

Potencial de Análisis Multidimensional: El uso de múltiples características de criminalidad (Criminality, Criminal markets, Criminal actors) permitió un análisis más rico y detallado, revelando las diversas dimensiones de la criminalidad que pueden influir en los patrones observados en los clusters.

REFERENCIAS

Global Organized Crime Index (n.d.). Retrieved from https://ocindex.net/rankings?f=rankings&view=List&group=Country

Ramírez, L. (2023, January 5). Algoritmo k-means: ¿Qué es y cómo funciona? Thinking for Innovation. https://www.iebschool.com/blog/algoritmo-k-means-que-es-y-como-funciona-big-data/

GeoPandas 0.14.4 — GeoPandas 0.14.4+0.g60c9773.dirty documentation. (n.d.). https://geopandas.org/en/stable/

Sanz, F. (2023, March 22). Algoritmo K-Means Clustering – aplicaciones y desventajas. The Machine Learners. https://www.themachinelearners.com/k-means/