PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

FACULTAD DE INGENIERÍA

ESCUELA DE INGENIERÍA INFORMÁTICA

Laboratorio 2 Ingeniería de negocios

Valentina San Martín Pacheco

Andrés Marcelo Vidal Soto

Profesor: Francisco González

Carrera: Ingeniería Civil Informática

JUNIO, 2024

1. Descripción del conjunto de datos y su contexto

Para el siguiente Laboratorio se escogió un dataset de Our World In Data distinto al usado en el Laboratorio 1 para poder graficar más fácilmente las variables, al ser las de este dataset de tipo cuantitativo. Este dataset muestra la relación entre la esperanza de vida promedio (Average Life Expectancy) y la cantidad de años que se espera que una persona viva con una pérdida de salud (enfermedad o discapacidad) ya sea a corto o a largo plazo (Expected years lived with disability or disease). Este dataset incluye además de las variables mencionadas previamente, el año de cada registro (entre 1990 y 2016), el país, el continente, población (estimado histórico) y un código del país.

Este laboratorio busca encontrar una tendencia entre la esperanza de vida promedio (Average Life Expectancy) y la cantidad de años que se espera que una persona viva con una enfermedad o discapacidad y compararlo con los continentes para evaluar la posible relación entre estos. Para esto se utilizará el algoritmo de clusterización K-Means.

2. Implementación

Para el desarrollo del presente laboratorio se siguieron los siguientes pasos:

1. Importación y procesamiento de datos:
    - Se importan las librerías necesarias: pandas, numpy, matplotlib.pyplot, seaborn, entre otras.
    - Se carga el dataset en chunks debido a su tamaño.
2. Preprocesamiento de Datos:
    - Se separan las columnas numéricas y no numéricas.
    - Se imputan los datos con valor NaN en columnas numéricas con la media y en columnas no numéricas con "no especificado".
    - Se eliminan las filas que no posean las columnas 'Years lived with disability' y 'Life Expectancy (IHME)' ya que no serán utilizadas para el análisis.
    - Se combinan las columnas nuevamente en un único dataframe imputado.
    - Se escalan los datos para dejarlos con una distribución normal y así disminuir la varianza para facilitar la visualización de los datos al graficarlos.
    - Se seleccionan las columnas de interés para el análisis: "Years lived with disability" y "Life Expectancy (IHME)".
3. Exploración de los datos:
    - Visualización de los primeros registros, información general y estadísticas descriptivas.
    - Creación de un gráfico de dispersión para visualizar la relación entre "Años vividos con discapacidad" y "Esperanza de vida".
    - Gráficos Boxplot por continente que muestran la esperanza de vida por continente y los años vividos con discapacidad por continente.
4. Análisis de Patrones con K-Means:

- Se busca el número de clusters óptimo utilizando la regla del codo en el gráfico inercia v/s número de clusters, donde el número resultante fue K = 3.
- Se aplica el algoritmo K-means con 3 clusters.
- Se añaden los resultados de la clusterización al dataframe original.
- Se visualizan los clusters en un gráfico de dispersión.
- Se calculan e imprimen las métricas de inercia, puntaje silhouette y puntaje Calinski-Harabasz.

5. Interpretación de resultados:
   - Se filtran los datos para excluir registros donde el continente no esté especificado.
   - Se crean gráficos de dispersión diferenciando por continente.

2.1. Técnica de Minería de Datos Utilizada: K-Means

La técnica de minería de datos utilizada en este Laboratorio corresponde al algoritmo de clusterización K-Means

2.1.1. Algoritmo K-Means

El algoritmo K-Means es un método de clustering que agrupa datos en clusters basándose en la similitud de características, funciona de la siguiente manera:

1. **Inicialización**: Se seleccionan aleatoriamente k centroides iniciales, donde k es el número de clusters que se desea obtener.
2. **Asignación**: Para cada punto de datos, se calcula la distancia (generalmente la distancia euclidiana) hacia cada centroide y se asigna al cluster cuyo centroide está más cercano.
3. **Actualización**: Se recalculan los centroides de los clusters como el promedio de todos los puntos asignados a ese cluster (el "centro de masa " del clúster).
4. **Repetición**: Los pasos 2 y 3 se repiten hasta que los centroides no cambien o se alcance un número máximo de iteraciones, ya que en este punto se asegura la convergencia en el algoritmo.

2.1.2. Medidas de Validación de Clusters

Estas medidas se utilizan para evaluar la calidad y coherencia de los clusters obtenidos

1. **Inercia (Within-Cluster Sum of Squares)**: Es la suma de las distancias cuadradas de cada punto de datos en un cluster a su centroide más cercano.

   Se calcula como:

   i. Inercia = $\sum_{j=1}^{K} \sum_{i=1} (\mathrm{N}(x_i^k - C_k)^2)$

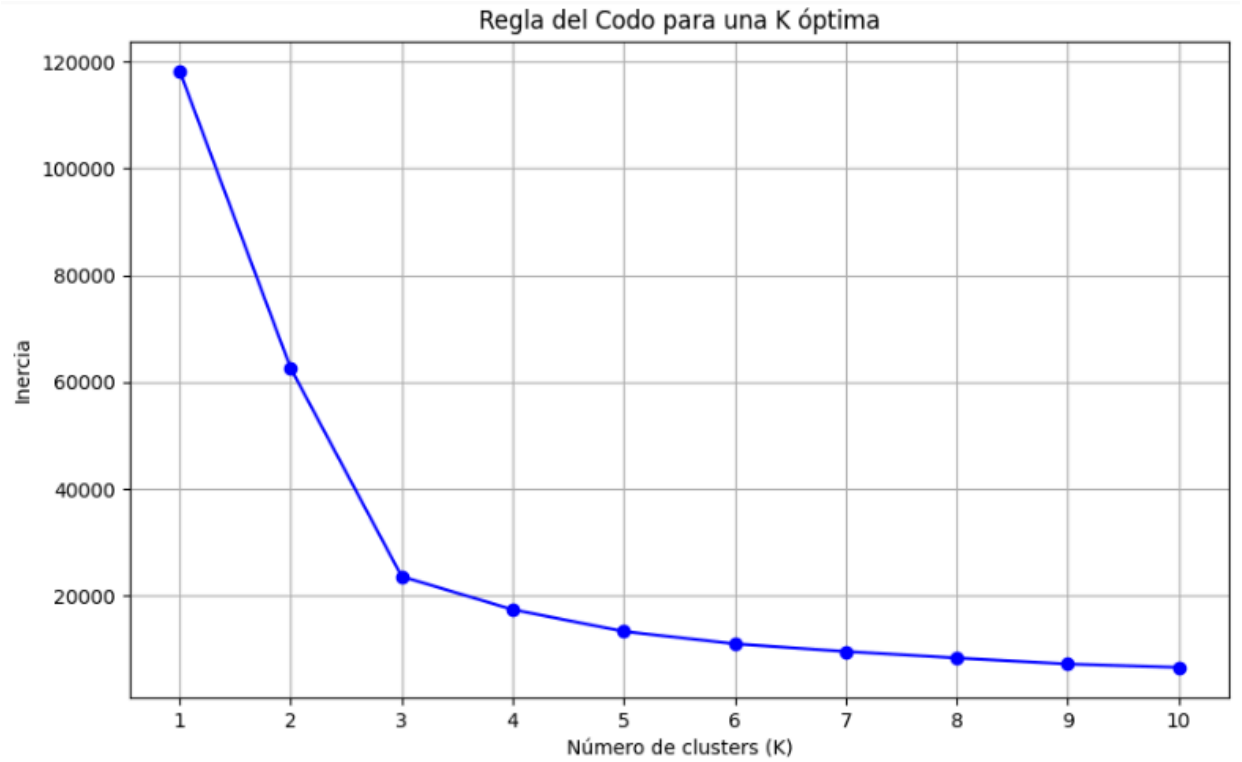
- donde:
  - K: número de clusters a generar
  - x_i^k : i-ésima instancia que pertenece al cluster k

- C_k: Centroide del cluster k
- N: Número total de instancias del cluster k

por lo que dada la diferencia $(x_i^k - C_k)^2$, que representa la distancia euclidiana entre un punto x_i^k y el centroide C_k de cada cluster k, mientras más cerca están las instancias

del centroide de su clúster, el valor de inercia será más bajo y por ende inercias más

pequeñas, serán inercias más atractivas. Ahora como naturalmente mientras más clusters se tengan, menor será la inercia pero no significa que el clustering sea mejor, por lo que el número de clusters se determina con la "regla del codo" que es el punto más minimizado (el más cercano al punto (0, 0)) que en este lab correspondería al punto con **k = 3** con un valor de Inercia del 23.643,007 (este gráfico aparece hecho previo a la implementación del K-Means para saber con qué k inicializar el algoritmo).



2. **Silhouette**:
   - Mide cuán similar es un punto a su propio cluster en comparación con otros clusters.
   - Varía entre -1 y 1, donde valores cercanos a 1 indican que los puntos están bien agrupados, mientras que valores cercanos a -1 indican que pueden haber sido asignados al cluster incorrecto.
   - Se calcula para cada punto como:
     i. $s(x) = \frac{b(x) - a(x)}{max\{a(x), b(x)\}}$

donde a(x) es la distancia promedio al resto de puntos en el mismo cluster, y b(x) es la distancia promedio al resto de puntos en el cluster más cercano diferente.

- En este laboratorio el resultado del silhouette score con K= 3 corresponde a 0,935, lo que indicaría que al ser más cercano a 1 que a -1, los puntos están bien agrupados.

3. **Calinski-Harabasz (Criterio de relación de varianza)**:
   - Es una medida que compara la dispersión dentro de los clusters con la dispersión entre los clusters.
   - Cuanto mayor sea el valor de Calinski-Harabasz, mejor será la estructura de clustering.
   - Se calcula como:
     i. $CH = \frac{B(k)}{W(k)} \text{ X } \frac{N-k}{k-1}$

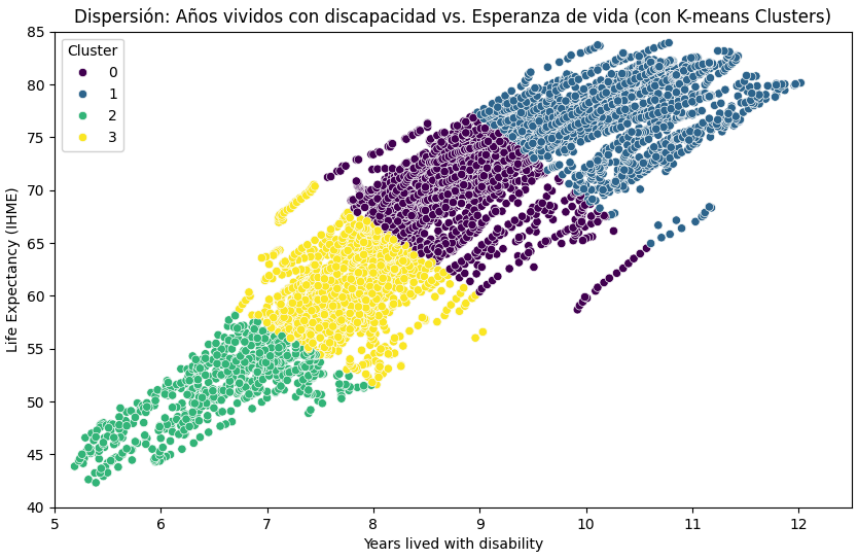
donde B(k) es la dispersión entre clusters, W(k) es la dispersión dentro de clusters, N es el número total de puntos, y k es el número de clusters.

- En este laboratorio con K= 3 el resultado de la medida de Calinski-Harabasz corresponde a 118.220,208, indicando una buena estructura de clustering para el dataset.

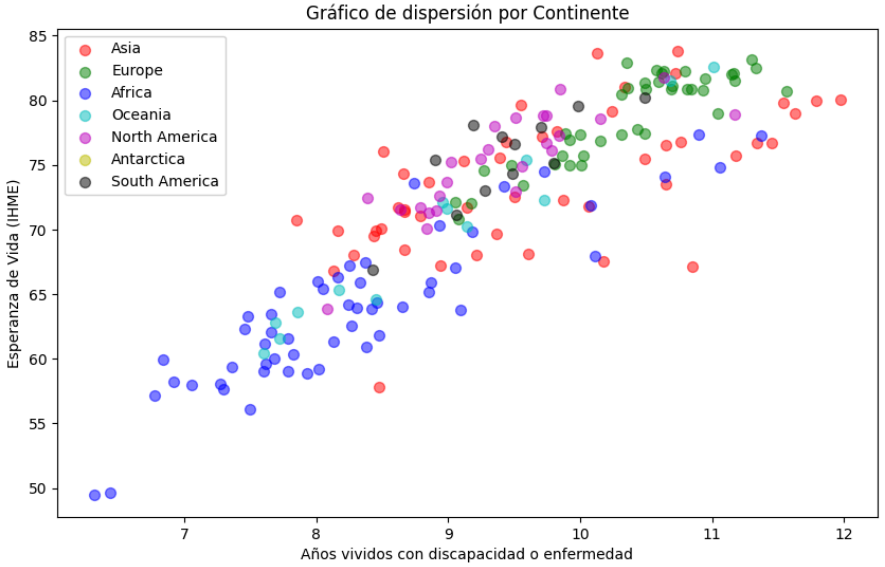
Estas medidas son útiles para seleccionar el número óptimo de clusters (usando métodos como el "codo" en el gráfico de inercia o maximizando Silhouette y Calinski-Harabasz) y para evaluar la coherencia de los clusters obtenidos por el algoritmo K-Means.

3. Análisis de patrones

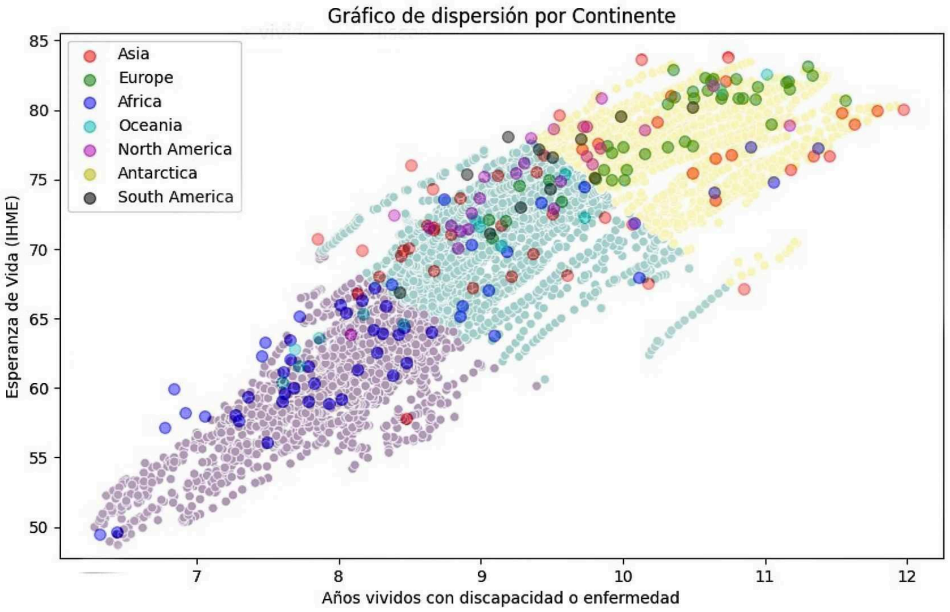
Para la visualización de los clusters identificados por el algoritmo K-Means se utilizó un gráfico de dispersión donde se visualiza la relación entre "Años vividos con discapacidad" y "Esperanza de vida", mostrando una tendencia general en los datos.



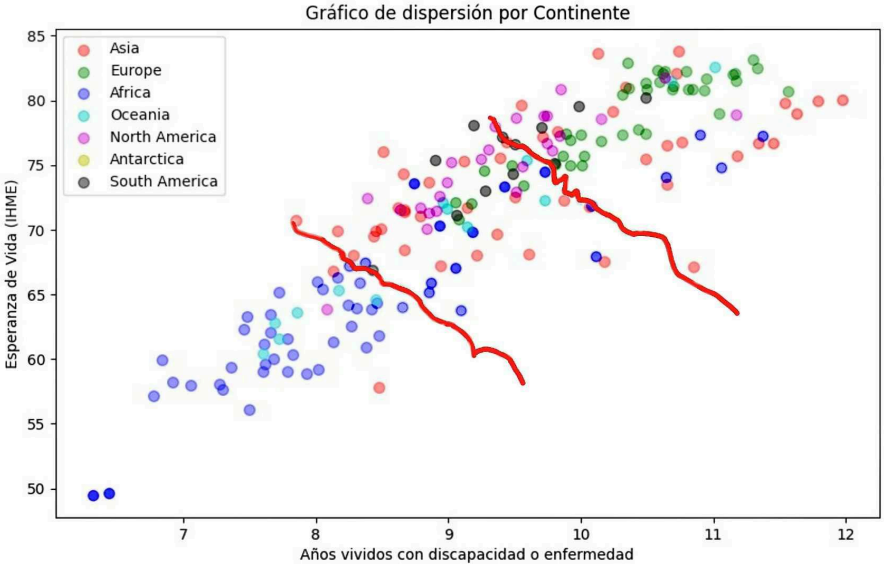
Luego, para evaluar la relación entre la esperanza de vida y los años vividos con discapacidad respecto a los continentes a los que pertenece cada instancia, se modeló un gráfico de dispersión por continente donde los datos se visualizan coloreados por continente en un gráfico de esperanza de vida v/s años vividos con discapacidad.



Ahora uniendo los dos gráficos para establecer la relación entre ambos según los clusters generados por el algoritmo quedaría así:

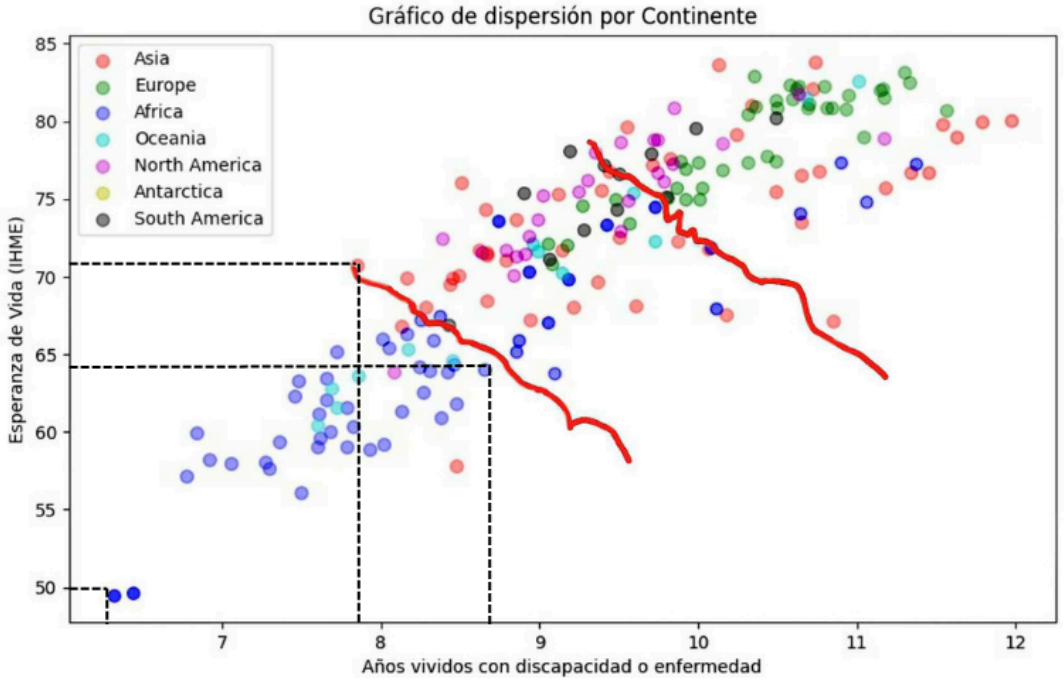


Luego, si marcamos las delimitaciones entre los clusters quedaría así:



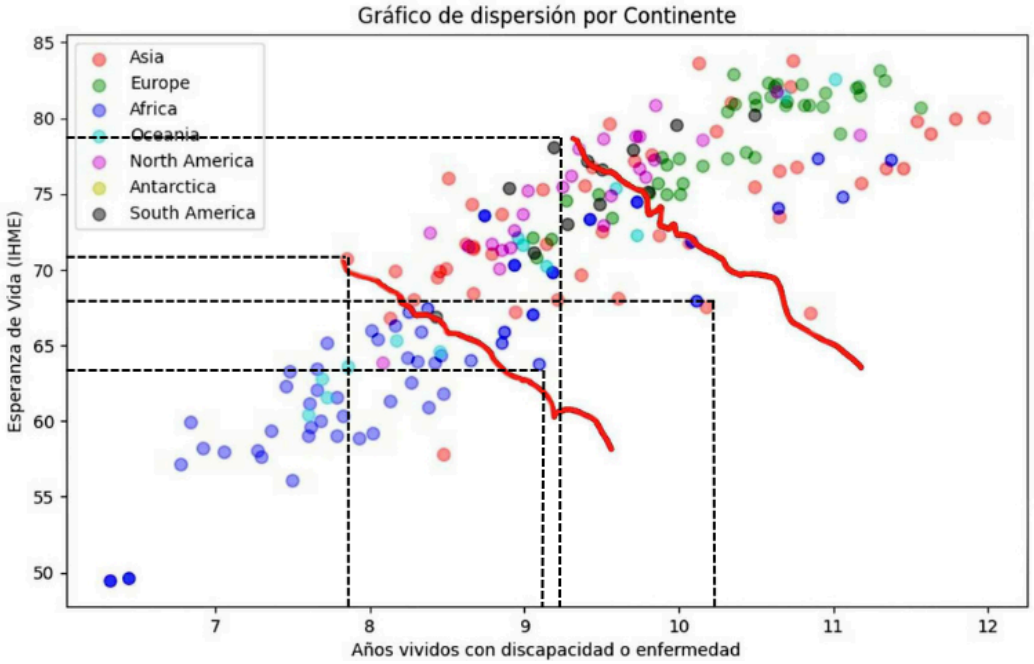
Resultando así más fácil de evaluar cada cluster por separado y establecer relaciones entre las variables “esperanza de vida”, “años vividos con discapacidad” y “continente”.

3.1. Cluster 1



Para el primer cluster, se puede ver que este está distribuido entre aproximadamente 6 años casi los 9 años de discapacidad, mientras que la esperanza de vida está distribuida entre los 50 años y los 70 años de edad. Además dentro de este se observa que la mayoría son países provenientes de África.

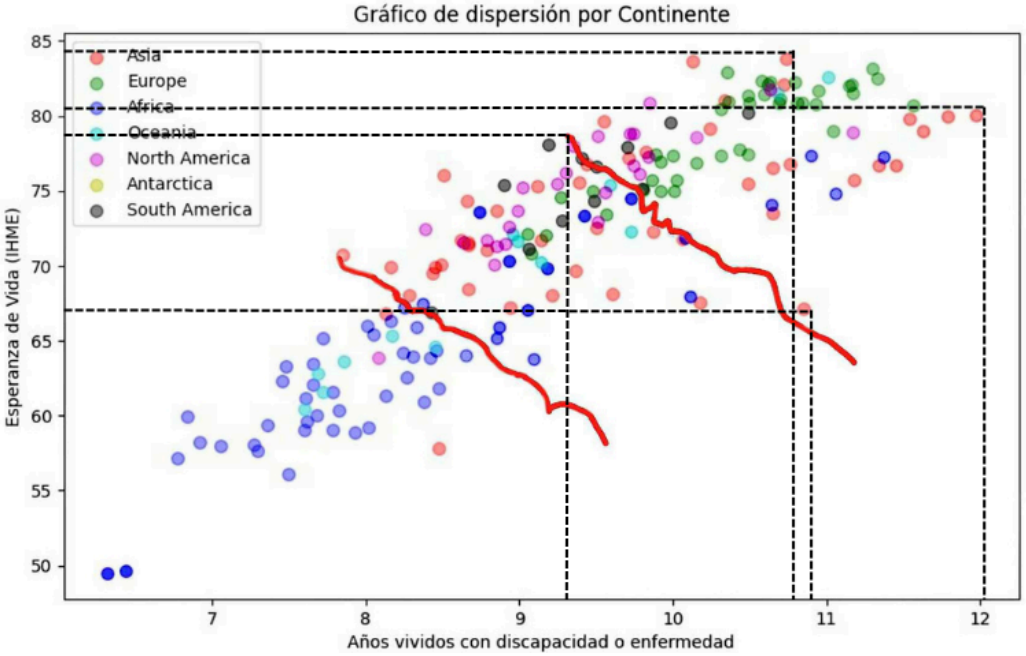
3.2. Cluster 2



Para el segundo cluster, se puede ver que este está distribuido entre aproximadamente 8 años y un poco más de 10 años de discapacidad, mientras que la esperanza de vida está distribuida entre casi 65 años hasta 80 años de edad aproximadamente.

Este es el cluster con mayor variedad de continentes pero si vemos la distribución de cada continente por cluster, en el cluster 2 están más de la mitad de los países de Norteamérica y de Sudamérica y también incluye países de Asia en una medida similar a como están también distribuidos en el cluster 3.

3.3. Cluster 3



Para el tercer cluster, se puede ver que este está distribuido entre un poco más de los 9 años y los 12 años de discapacidad, mientras que la esperanza de vida está distribuida entre un poco más de los 65 años hasta casi los 85 años de edad aproximadamente.

Respecto a los continentes, en este cluster se encuentra la mayoría de los países provenientes de Europa y comparte algunos países de Asia con el cluster 2.

3.4. Resumen de Análisis de Patrones

- Cluster 1:
    - Años con discapacidad: [6, 9]
    - Esperanza de Vida: [50, 70]
    - Continentes: {África}
- Cluster 2:
    - Años con discapacidad: [8, 10]
    - Esperanza de Vida: [65, 80]
    - Continentes: {Norteamérica, Sudamérica, Asia}

- Cluster 3:
    - Años con discapacidad: [9, 12]
    - Esperanza de Vida: [65, 85]
    - Continentes: {Europa, Asia}

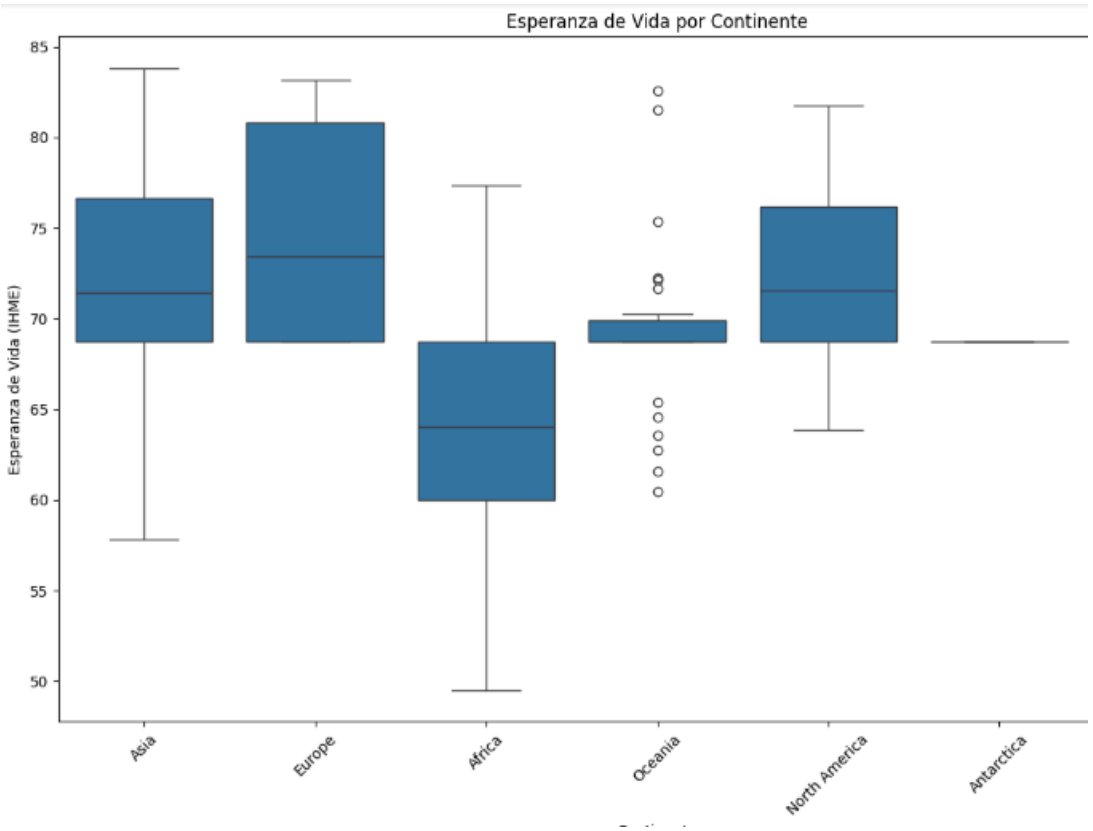
Ahora respecto al resto de continentes, Oceanía se encuentra distribuida entre los tres clusters, por lo que se requiere de los boxplots para identificar la poca tendencia que tiene, la cual correspondería a:

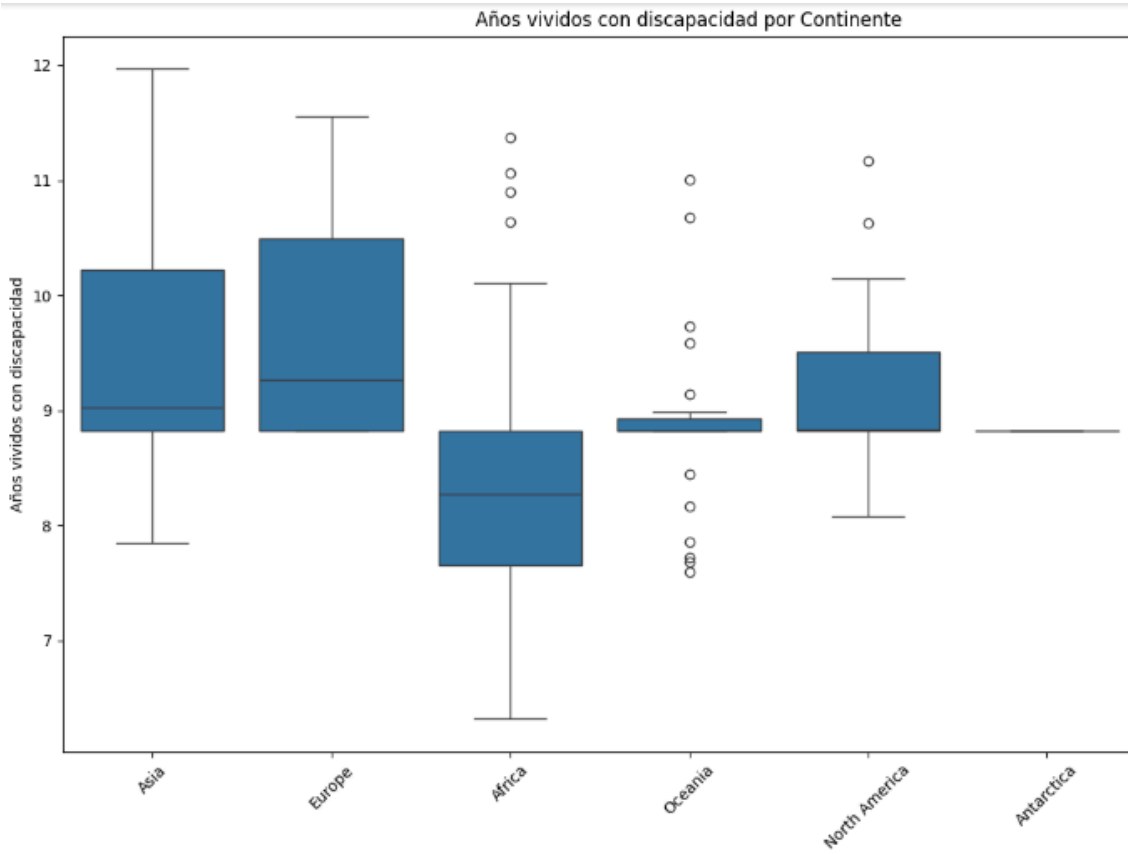
- Años con discapacidad → 9
- Esperanza de Vida → 70

Ahora respecto a Antártica, al tener tan pocos países no se visualiza en el gráfico de dispersión, por lo que también resulta necesario revisar el boxplot, del cual se observan las siguientes tendencias:

- Años con discapacidad → 9
- Esperanza de Vida → 70

Por lo que se podría inferir que el cluster al que más se asemejarían tanto Oceanía como Antártica sería al cluster 2.





4. Insights de los datos respecto a la toma de decisiones

Inferencias respecto a los clusters obtenidos de las instancias del dataset:

1. **Patrones de salud global:** La relación entre "Años vividos con discapacidad" y "Esperanza de vida" muestra variaciones significativas que pueden estar influenciadas por factores regionales y económicos.
2. **Clusters de salud:** La clusterización K-means revela grupos específicos de países con características similares, lo que puede ayudar a identificar patrones comunes y enfoques regionales de salud pública.
3. **Variación continental:** Las diferencias en los gráficos de dispersión por continente indican que las estrategias de salud pública podrían necesitar personalización para ser más efectivas en diferentes regiones.

Ayuda en la toma de decisiones:

Estas recomendaciones pueden guiar a los tomadores de decisiones en la implementación de políticas de salud más efectivas y en la asignación de recursos para mejorar los resultados de salud global.

1. **Políticas de salud personalizadas:** Los países y regiones deberían considerar políticas de salud específicas basadas en los patrones observados en los clusters y los análisis continentales.

2. **Investigación adicional:** Sería útil investigar más a fondo los factores subyacentes que contribuyen a las diferencias en "Años vividos con discapacidad" y "Esperanza de vida" entre los clusters y los continentes.
3. **Monitoreo continuo:** Implementar sistemas de monitoreo continuo para actualizar y ajustar las estrategias de salud pública basadas en datos recientes y patrones emergentes.