Algoritmos de Inteligência Artificial para clusterização - INFNET

Os dados abaixo também estão no arquivo do projeto, no formato Markdown.

1. **Infraestrutura**
   a. Criação e ativação de ambiente virtual com Virtualenv:

```
PS C:\Users\andre\projetos\clusterizacao_i_infnet> pyenv exec python -m venv .venv
PS C:\Users\andre\projetos\clusterizacao_i_infnet> .venv/Scripts/activate
(.venv) PS C:\Users\andre\projetos\clusterizacao_i_infnet>
```

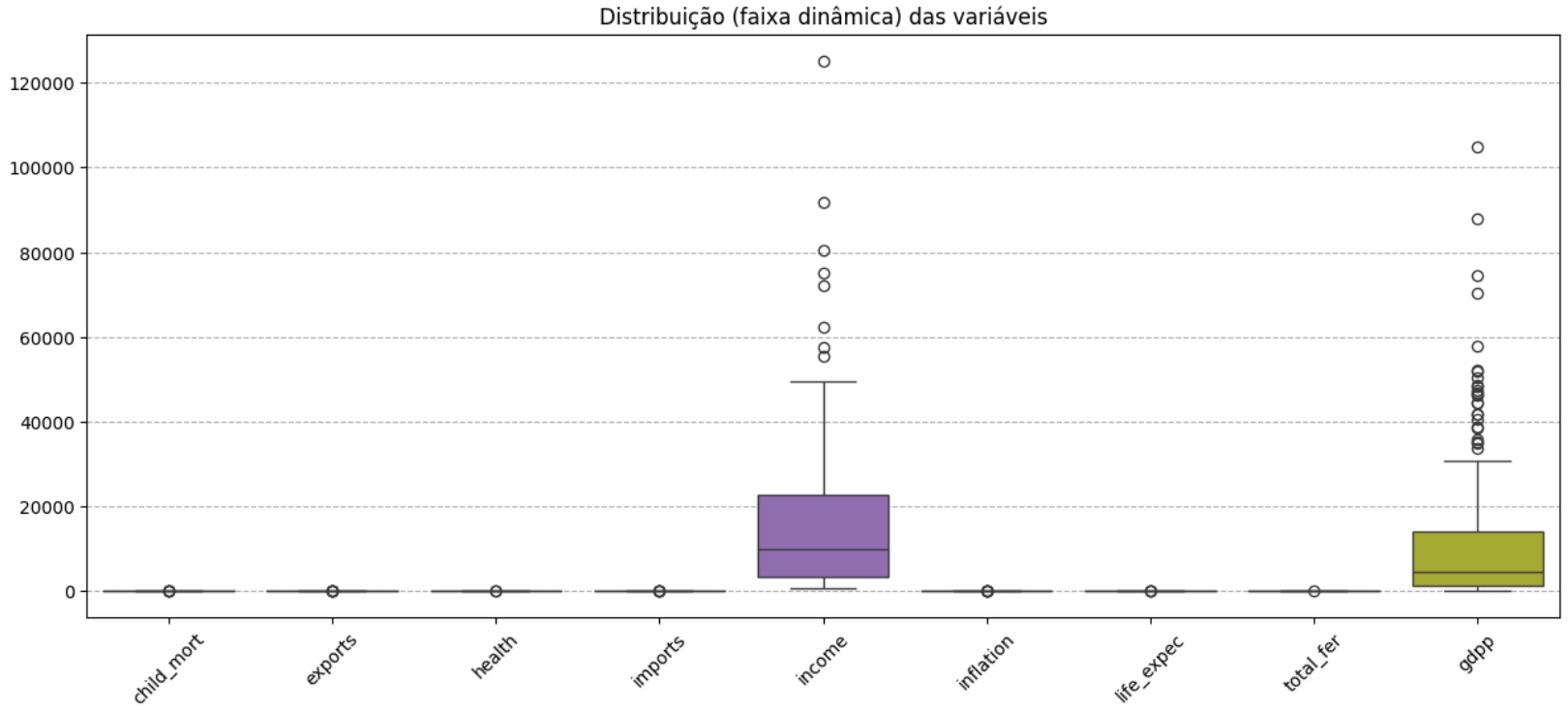
   b. Repositório git: https://github.com/andreiluizpereira/clusterizacao_i_infnet.git

2. **Escolha de base de dados**
   a. Fonte: https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data

   b. Quantos países existem no dataset?
   R: Existem 167 países

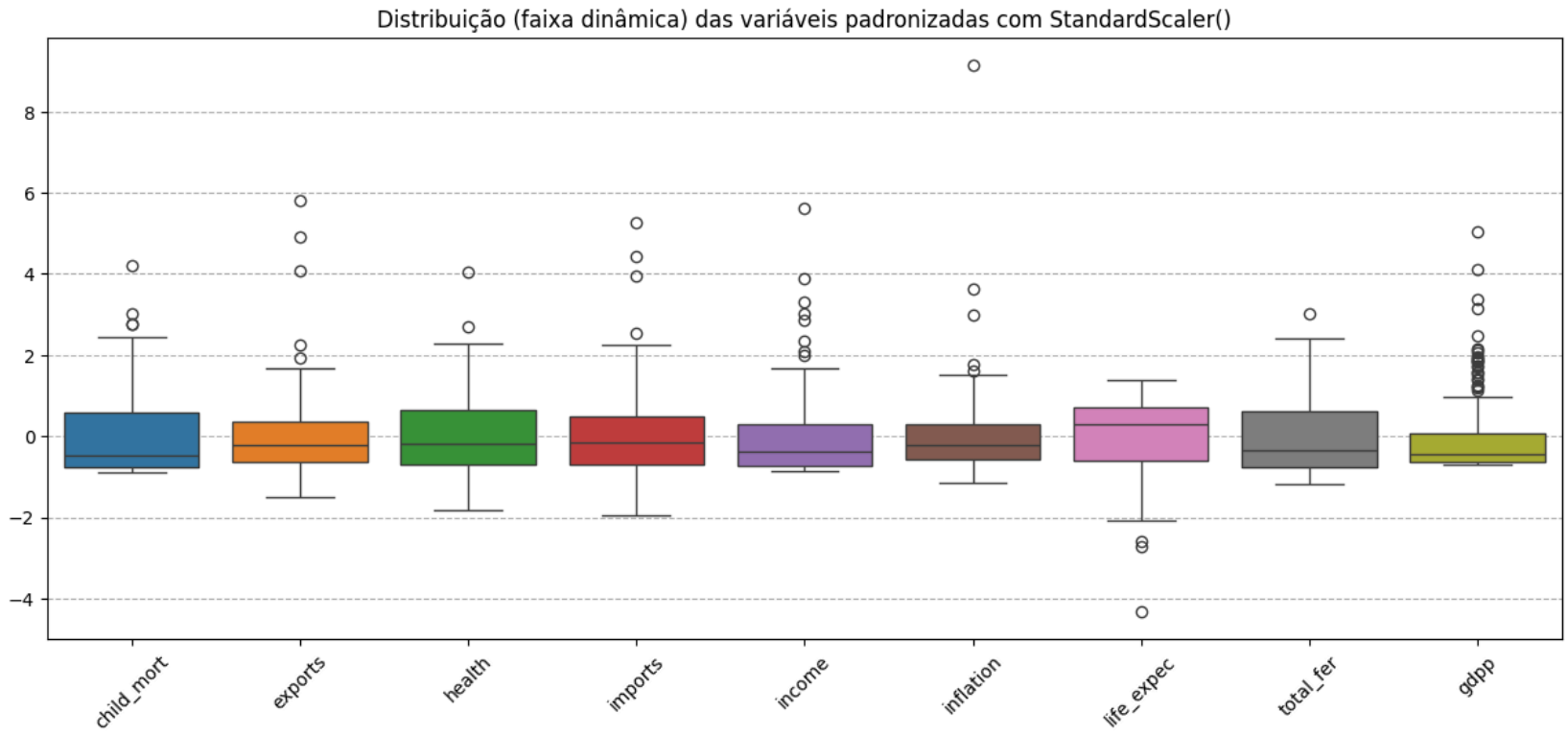
   c. Mostre através de gráficos a faixa dinâmica das variáveis que serão usadas nas tarefas de clusterização. Analise os resultados mostrados. O que deve ser feito com os dados antes da etapa de clusterização?

R: **Gráfico da faixa dinâmica das variáveis:**



Observando os boxplots gerados, nota-se que as variáveis apresentam escalas bastante distintas. Variáveis como **`gdpp`** e **`income`** possuem valores consideravelmente maiores em comparação às demais.
Esta diferença de escala representa um problema, pois algoritmos baseados em distância euclidiana, como o K-Means, atribuem maior peso às variáveis com valores absolutos maiores. Portanto, torna-se necessária a padronização dos dados antes de realizar a clusterização.

d. Realize o pré-processamento adequado dos dados.
R: **Dados normalizados:**



3. **Clusterização**
    - **a.** Para os resultados do K-Médias, interprete cada um dos clusters obtidos citando:
        - i. Qual a distribuição das dimensões em cada grupo:

R: Contagem de países por cluster (K-Means):

Cluster 0: 86 países

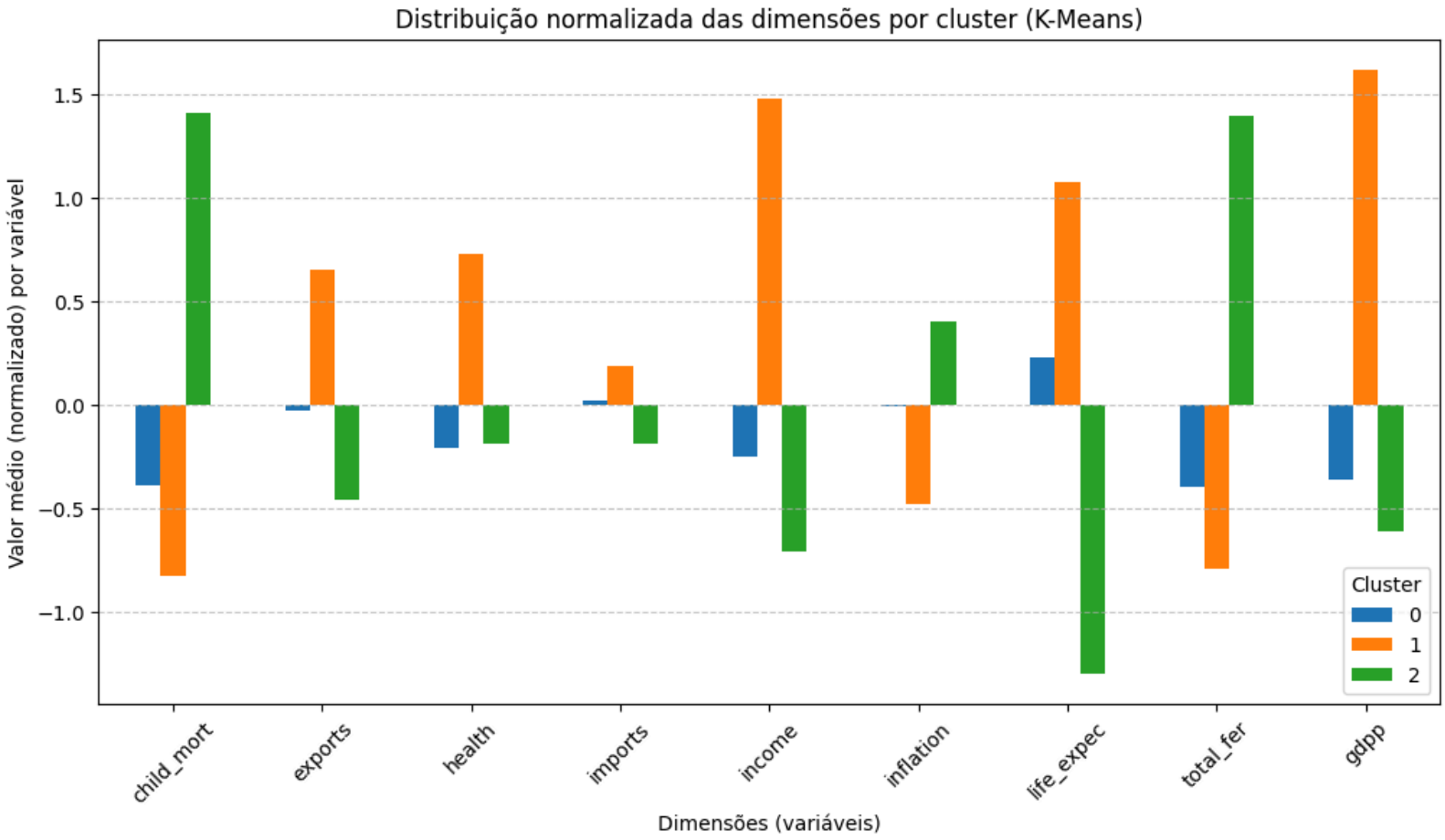
Cluster 1: 36 países

Cluster 2: 45 países

A partir da análise do gráfico de barras com as médias normalizadas, observa-se o seguinte perfil para cada cluster:
Cluster 0: Países com indicadores intermediários, situando-se entre os extremos de desenvolvimento econômico e social.

Cluster 1: Países desenvolvidos, caracterizados por elevados valores de `**gdpp**` e `**income**`, alta expectativa de vida (`**life_expec**`) e baixa mortalidade infantil (`**child_mort**`).

Cluster 2: Países em desenvolvimento, apresentando baixo PIB e renda per capita, alta mortalidade infantil e menor expectativa de vida.



ii. O país que, de acordo com o algoritmo, melhor representa o seu agrupamento. Justifique.

R:

Cluster 0:

País representativo: Suriname

Distância ao centróide: 0.7198

Cluster 1:
País representativo: Iceland
Distância ao centróide: 0.7318

Cluster 2:
País representativo: Guinea
Distância ao centróide: 0.7704

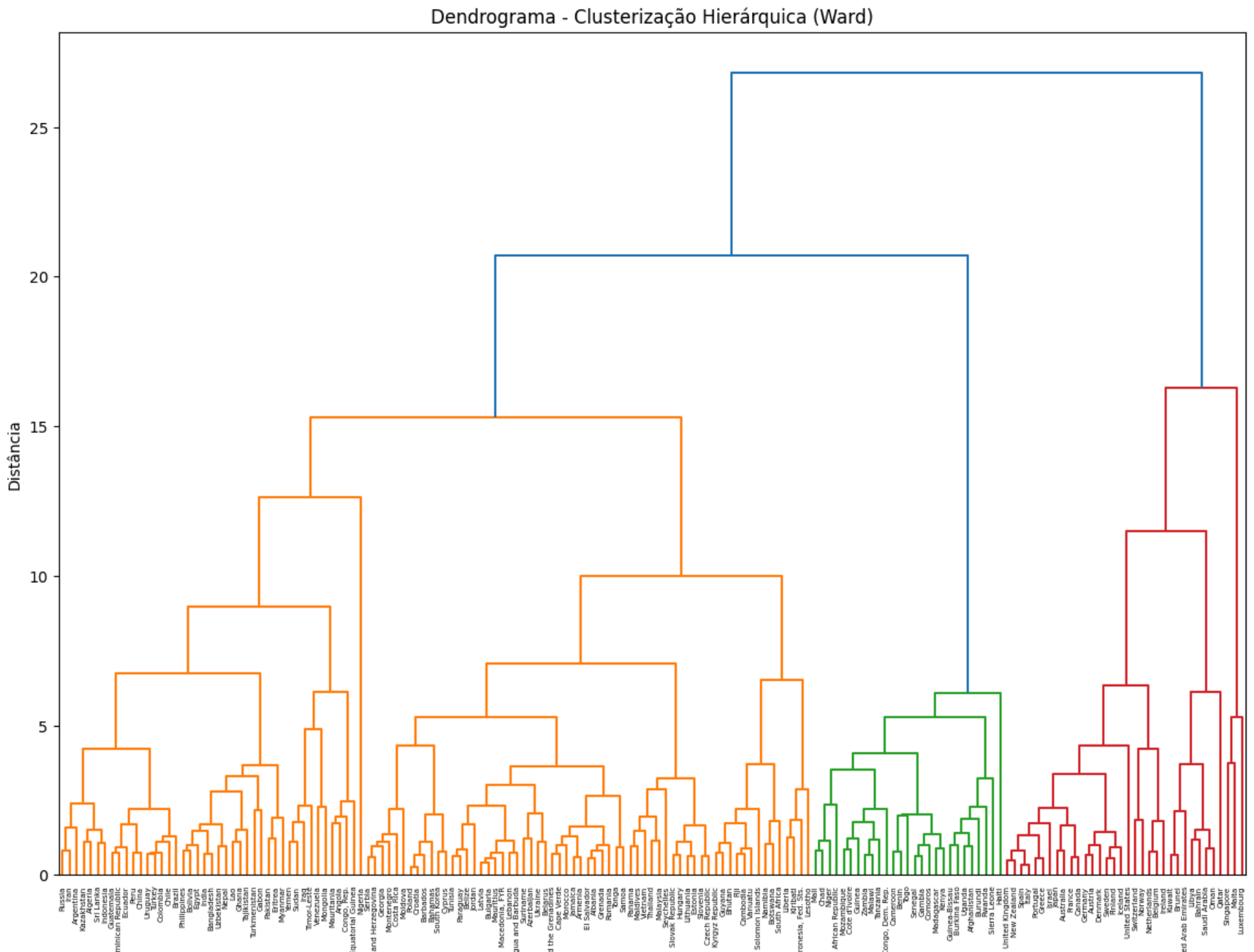
O país representativo de cada cluster corresponde àquele mais próximo do centróide, funcionando como o elemento mais típico do grupo.

Cluster 0 - Suriname: Apresenta valores intermediários nas variáveis analisadas, posicionando-se entre os dois extremos socioeconômicos identificados.

Cluster 1 - Iceland: País desenvolvido que exemplifica o perfil do cluster com alta renda per capita e elevada expectativa de vida.

Cluster 2 - Guinea: Representa o perfil de país em desenvolvimento, com baixa renda e alta taxa de mortalidade infantil.

b. Para os resultados da Clusterização Hierárquica, apresente o dendrograma e interprete os resultados.


Dendrograma - Clusterização Hierárquica (Ward)
Distância
0
5
10
15
20
25

O dendrograma apresenta a estrutura hierárquica de agrupamento dos países. A altura das ligações indica a distância entre os clusters - quanto maior a altura, maior a dissimilaridade entre os grupos sendo unidos.

É possível identificar a formação de três grandes grupos, resultado consistente com o obtido pelo K-Means. Países com perfis semelhantes (como Noruega e Suíça) são agrupados nos níveis iniciais da hierarquia, enquanto países com características muito distintas apenas se unem nos níveis superiores.

O método Ward busca minimizar a variância entre os clusters, resultando em grupos com tamanhos relativamente equilibrados.

c. Compare os dois resultados, aponte as semelhanças e diferenças e interprete.
R:

Semelhanças:
Ambos os algoritmos identificaram três grupos principais com perfis socioeconômicos distintos: países desenvolvidos, em desenvolvimento e subdesenvolvidos.
A composição central dos clusters é bastante similar, com países como Suíça, Noruega e Luxemburgo consistentemente agrupados juntos em ambos os métodos.

Diferenças:
O K-Means requer a definição prévia do número de clusters, enquanto a clusterização hierárquica permite esta escolha a posteriori através da análise do dendrograma.
O K-Means apresenta variabilidade nos resultados devido à inicialização aleatória dos centroides, ao passo que o método hierárquico produz resultados determinísticos.
O dendrograma gerado pelo método hierárquico oferece informações adicionais sobre a estrutura e relações entre os grupos.

Interpretação:
Ambos os métodos demonstraram eficácia na segmentação dos países. O K-Means destaca-se pela eficiência computacional e simplicidade de implementação. Já a clusterização hierárquica fornece uma visão mais detalhada da estrutura dos grupos.

4. **Escolha de algoritmos**

R:

a. Escreva em tópicos as etapas do algoritmo de K-médias até sua convergência.

Inicialização: Definir o número K de clusters e inicializar K centróides de forma aleatória no espaço de dados.
Atribuição: Calcular a distância de cada ponto para todos os centróides e atribuí-lo ao cluster correspondente ao centróide mais próximo.
Atualização: Recalcular a posição de cada centróide como a média aritmética de todos os pontos pertencentes ao respectivo cluster.
Convergência: Repetir os passos de atribuição e atualização até que os centróides se estabilizem ou até atingir o número máximo de iterações estabelecido.

b. O algoritmo de K-médias é sensível a outliers nos dados. Explique.

A sensibilidade do K-Means a outliers decorre do cálculo do centróide pela média aritmética dos pontos. Quando há presença de outliers (pontos com valores distantes do conjunto principal), estes exercem influência desproporcional sobre a média, deslocando o centróide para longe da região de maior concentração de pontos do cluster.
Esta distorção compromete a representatividade do centróide e pode resultar em classificações incorretas.

c. Por que o algoritmo de DBScan é mais robusto à presença de outliers?

O DBSCAN fundamenta-se no conceito de densidade de pontos, classificando-os em três categorias:
Pontos Centrais: Possuem quantidade mínima de vizinhos (`**min_samples**`) dentro de um raio definido (`**eps**`).
Pontos de Borda: Encontram-se na vizinhança de pontos centrais, mas não atendem ao critério de ponto central.
Ruído: Pontos isolados que não se enquadram nas categorias anteriores.
Como outliers caracterizam-se por estarem em regiões de baixa densidade, o DBSCAN os identifica naturalmente como ruído, impedindo que influenciem a formação dos clusters, que são construídos exclusivamente a partir de regiões densas.