Agrupamento de países por indicadores de felicidade utilizando k-means (World Happiness Report).

Antonio V. O. Borssato

Departamento de Informática – Universidade Federal do Espírito Santo (UFES)
Av. Fernando Ferrari, 514 – Goiabeiras – Vitória – ES – Brasil
CEP: 29075-910

(antonio.borssato@edu.ufes.br)

Abstract*: Happiness is one of the key socio-economic indicators in a society. In this context, this study aims to cluster country data from the World Happiness Report (WHR) between the years 2015 and 2019 using the unsupervised K-means algorithm, with the goal of identifying groups of countries that share similar characteristics based on happiness indicators. To achieve this, a quantitative and qualitative analysis was conducted, including the aggregation of WHR data, the calculation of descriptive statistics to better understand the dataset, and the application of K-means to recognize patterns among the selected countries.*

Resumo: *A felicidade é um dos principais indicadores socioeconômicos em uma sociedade. Diante disso, este trabalho tem como objetivo agrupar dados de países do World Happiness Report (WHR) entre os anos 2015 e 2019 através do algoritmo não supervisionado K-means, buscando observar grupos de países que compartilham características semelhantes baseadas nos indicadores de felicidade. Para isso, foi realizada uma pesquisa quantitativa e qualitativa, agrupando informações do WHR, calculando estatísticas descritivas para entender os dados e aplicando o K-means para o reconhecimento de padrões nos países selecionados.*

1. Introdução

A felicidade, enquanto indicador de bem-estar social e econômico, tem ganhado destaque nas últimas décadas como métrica fundamental para avaliação do desenvolvimento humano. Em escala mundial, o *World Happiness Report* é um relatório que consolida dados de diversos países em indicadores como PIB *per capita*, suporte social, expectativa de vida, liberdade, generosidade e percepção de corrupção [*World Happiness Report*, 2025], oferecendo uma base robusta para análises comparativas entre diferentes nações. Esses indicadores não somente refletem a qualidade de vida das populações, mas também servem como subsídio para políticas públicas e estratégias socioeconômicas globais.

Neste contexto, a análise não supervisionada surge como uma opção para identificar padrões intrínsecos nos dados, permitindo agrupar países com características semelhantes sem rótulos prévios. O algoritmo *K-means* é um dos métodos de agrupamento mais populares usados em aprendizado de máquina, onde o agrupamento estipula que um ponto de dados pode existir em apenas um *cluster* [IBM, 2024]. Amplamente utilizado em ciência de dados, o *K-means* emprega o conceito de centroides como protótipos representativos dos grupos. Esse centroide simboliza o centro do grupo, sendo calculado pela média de todos os objetos presentes no agrupamento. [Fontana e Naldi, 2009]

Diante disso, o principal objetivo do trabalho é a aplicação do *K-means* ao WHR nos anos 2015 a 2019, buscando agrupar os países em que o indicador de felicidade se relaciona. Para isso, foi realizada uma pesquisa quantitativa e qualitativa extraindo dados do próprio

WHR, calculando estatísticas descritivas para compreender as características dos dados e aplicando o *K-means* para reconhecer padrões dos agrupamentos de países.

2. Definição do Problema

Apesar do relatório anual WHR ter aumentado a disponibilidade de informações sobre o bem-estar global, a visão comumente retirada desse documento se restringe às análises estáticas, muitas vezes baseadas no ranking dos países. A identificação de padrões não óbvios na distribuição da felicidade, por outro lado, ainda é um desafio complexo, uma vez que explorar relações multidimensionais entre indicadores socioeconômicos não é um trabalho trivial.

Este trabalho busca resolver três questões principais: como os países se distribuem geográfica e culturalmente em relação à felicidade; quais indicadores são determinantes para diferenciar os clusters; e se há correlação entre desenvolvimento econômico (países desenvolvidos vs. em desenvolvimento) e a classificação nos grupos identificados.

3. Descrição do Conjunto de dados

Os dados foram coletados através do Kaggle. Trata-se de uma plataforma online criada em 2010 para hospedar competições de ciência de dados [Vassalo, 2021], e também agrupa diversos tipos de banco de dados para análise, incluindo o WHR.

As principais variáveis são:

- *Country*: nome do país.
- *Region*: Região do país.
- *Economy* (GDP per Capita): PIB per capita.
- *Family*: Suporte social percebido.
- *Health* (Life Expectancy): Expectativa de vida saudável.
- *Freedom*: Percepção de liberdade para escolhas de vida.
- *Generosity*: Nível de generosidade reportado.
- *Trust (Government Corruption)*: Percepção de confiança governamental.
- *Happiness Score e Happiness Rank*: Métricas de felicidade agregada.
- *Dystopia Residual*: Distopia se refere a um país hipotético que possui todas as piores métricas nas 6 categorias deste dataset. Ele foi criado como ponto de comparação nos dados. O Dystopia Residual é o valor do Happiness Score da Distopia + o valor residual ou o valor não explicado para cada país.
- *Standard Error*: erro padrão do *Happiness Score.*

Todas as variáveis são quantitativas contínuas, exceto *Country* e *Region*, que são qualitativas nominais.

4. Metodologia

A metodologia deste trabalho foi dividida em três etapas fundamentais: coleta e pré -processamento de dados; análise descritiva e exploratória de dados; clusterização com K-means. Todos os procedimentos foram realizados para cada ano e também para a média geral dos dados entre 2015 e 2019, entretanto, essa seção mostrará apenas os resultados gerais de 2015 a 2019. O teste completo encontra-se no notebook.

4.1. Coleta e pré-processamento

Na etapa de coleta e pré-processamento, foram extraídos dados do *World Happiness Report* do Kaggle (2015–2019). Inicialmente, os nomes das colunas foram padronizados para manter a consistência. A coluna "*Region*" passou por um mapeamento de países para regiões (baseado em 2015 e 2016, que possuíam essa coluna), com ajustes nas discrepâncias de nomenclatura e investigação dos países não mapeados. Em seguida, selecionou-se o "*Happiness Score*" e seis indicadores (*Economy, Family, Health, Freedom, Trust e Generosity*), removendo-se linhas com dados faltantes. Por fim, os dados foram filtrados para considerar apenas os países que estão presentes em todos os anos.

4.2. Análise descritiva e exploratória de dados

Na fase de análise descritiva e exploratória de dados, foram calculadas estatísticas descritivas (média, desvio padrão e quartis) para compreender as características e a distribuição dos dados e, posteriormente, é realizada uma análise de correlação (através de matrizes e *pairplots*) para identificar as relações entre os indicadores e o "*Happiness Score*", o que permite a identificação de tendências gerais e variabilidades ao longo do período analisado. Com isso, percebeu-se que o score de felicidade é positivamente correlacionado, em ordem decrescente, com Economia (BIP per capta), Família (Suporte social) e Saúde (Expectativa de vida). Avaliando cada uma dessas variáveis, sem considerar o score de felicidade, tem-se que:

- Economia (BIP per capta): maior correlação é positiva, com a Saúde (Expectativa de vida).
- Família (Suporte social): maior correlação é positiva, com Economia (BIP per capta).
- Saúde (Expectativa de vida): maior correlação é positiva, com Economia (BIP per capta).

Isso indica que países mais ricos, com melhor expectativa de vida e laços familiares mais fortes, tendem a ter maiores índices de felicidade. Esses fatores podem criar clusters distintos de países desenvolvidos e em desenvolvimento, portanto, isso será avaliado no decorrer do trabalho.

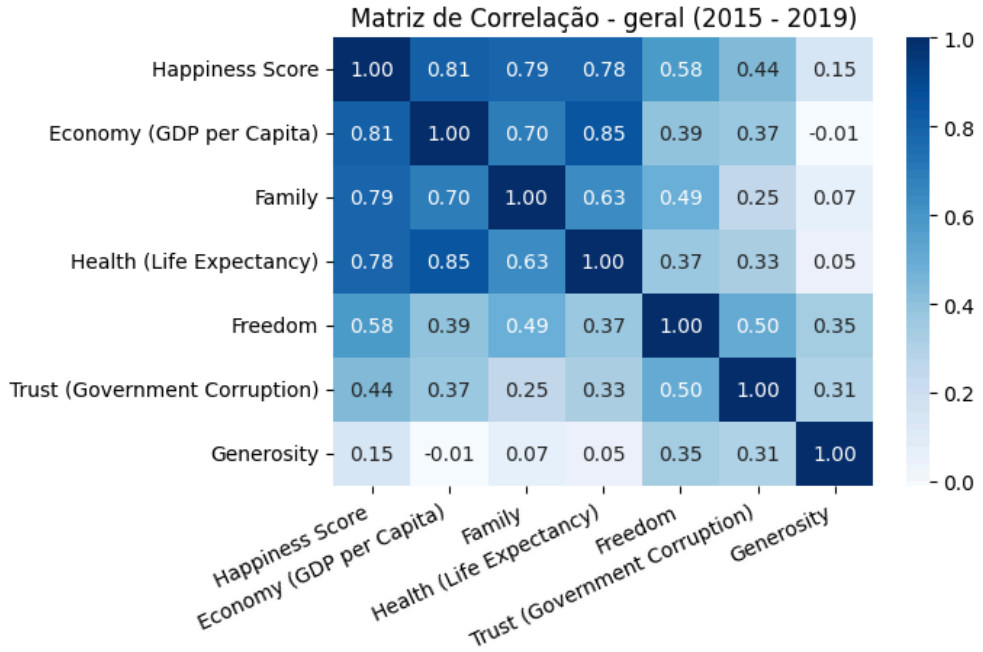

Figura 1: matriz de correlação dos dados gerais (2015-2019)

4.3. Clusterização com *K-Means*

Por fim, na clusterização com *K-Means* os dados foram normalizados com *StandardScaler* por conta da sensibilidade do algoritmo para distâncias. Diante da alta correlação entre variáveis como *Economy* e *Health*, e da baixa correlação entre *Generosity* e *Happiness Score*, a técnica de redução de dimensionalidade (SVD) foi avaliada para averiguar o rank efetivo da matriz.

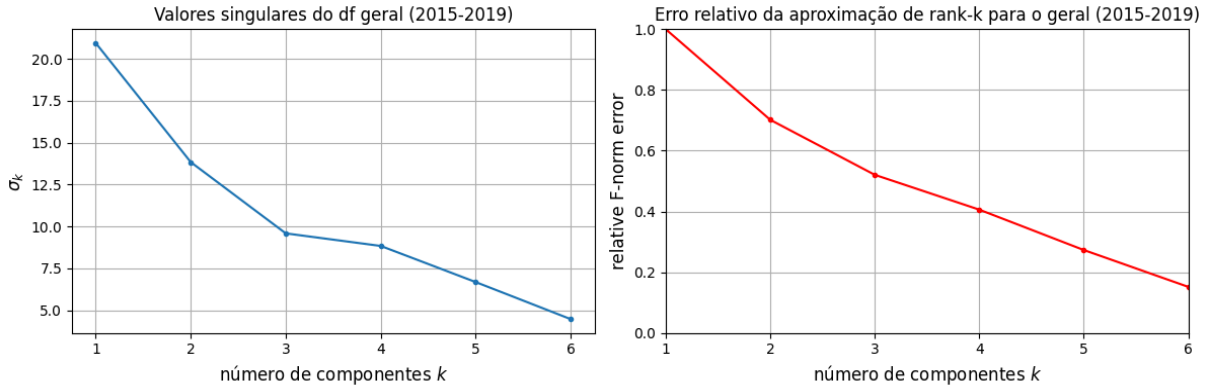

Figura 2: valores singulares e erro de aproximação do SVD

Percebe-se que os dados não apresentam sinais de baixo rank efetivo. A redução de uma única dimensão representa a perda de aproximadamente 30% de informação do conjunto. Nesse caso, não foi realizada a redução de dimensionalidade.

Em seguida, o número ótimo de clusters foi definido pela análise coletiva do método *Elbow* (Cotovelo), silhueta, Calinski-Harabasz e Davies-Bouldin.

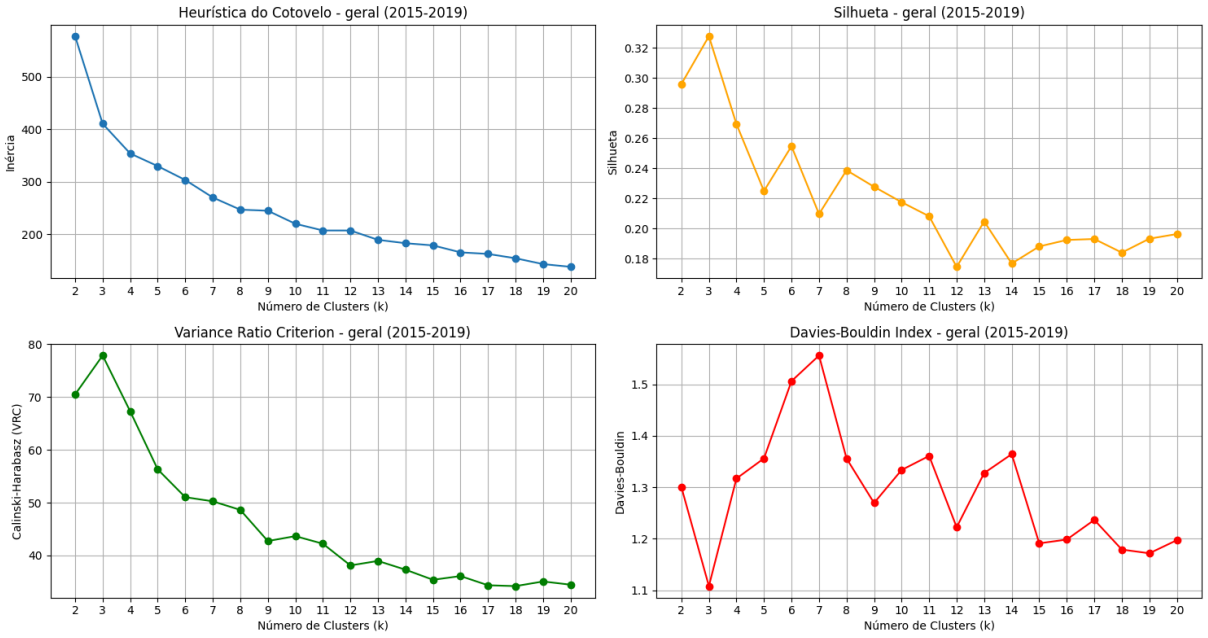

Figura 3: métricas de avaliação do hiperparâmetro k.

Como foi observado que as métricas apontavam para k=3 em todos os anos (picos na silhueta e variance ratio criterion, desaceleração da inércia na heurística do cotovelo e vale no Davies-Bouldin), o algoritmo *K-Means* foi aplicado aos dados com esse valor de clusters. Os *labels* resultantes foram salvos com as variáveis separadas para análise.

5. Resultados

Após o agrupamento, os resultados apresentaram separação clara e consistente de clusters para todos os anos, observada na boa separação dos grupos no mapa, uma matriz de distâncias consistente e o gráfico de dispersão das 3 variáveis mais importantes.

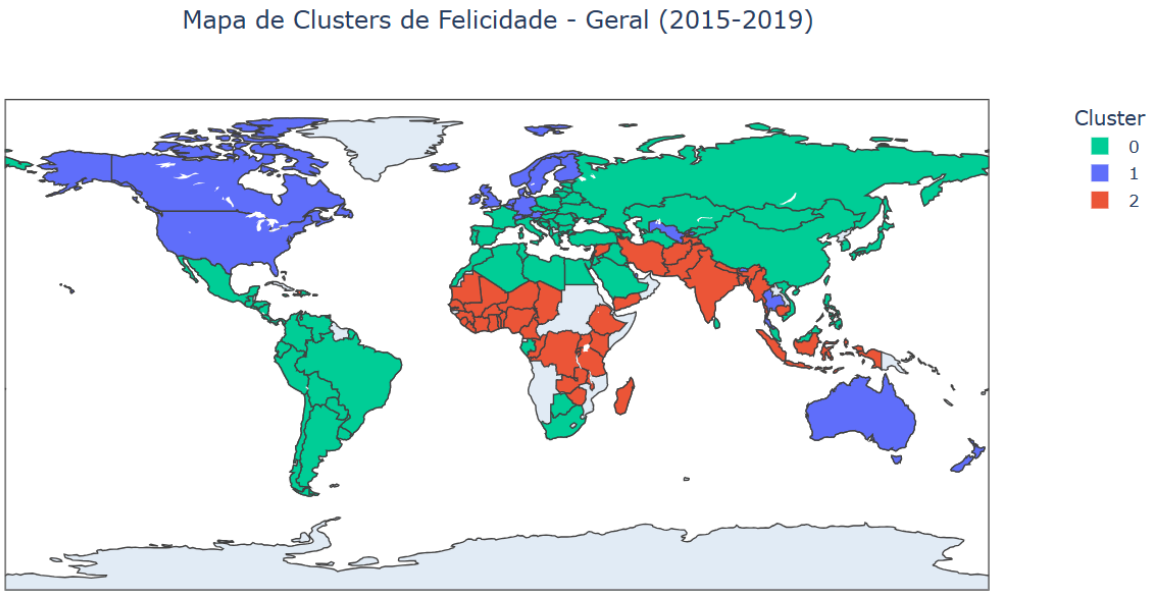

Figura 4: mapa com a separação dos cluster destacando as regiões (2015-2019)

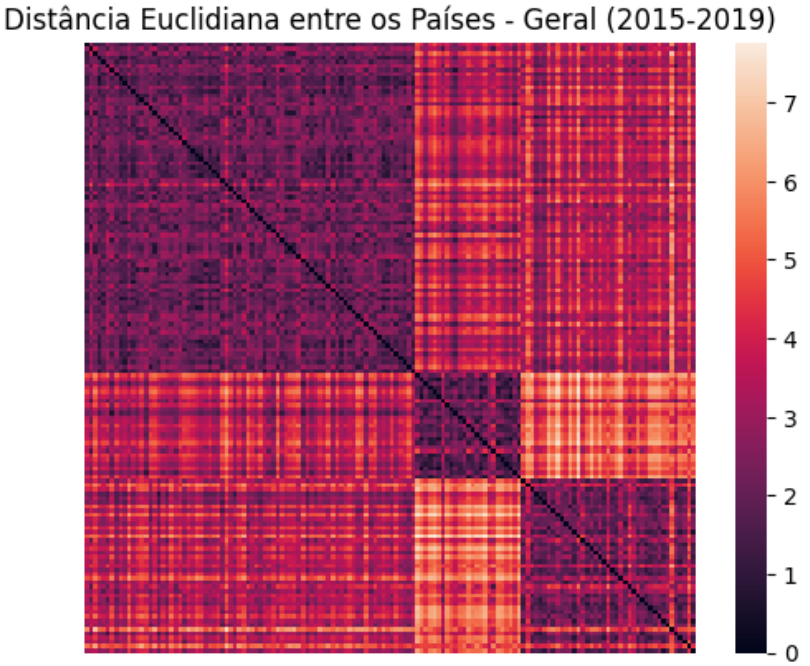

Figura 5: matriz de distâncias euclidianas, destacando 3 grupos distintos (2015-2019)

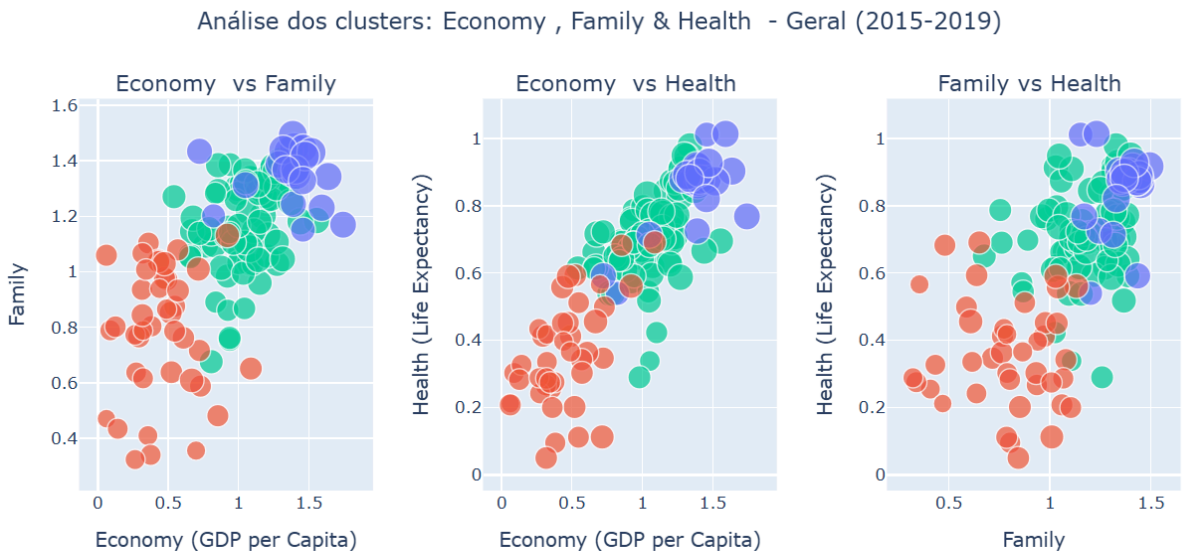

Figura 6: gráficos de dispersão dos clusters nas 3 variáveis mais importantes (2015-2019)

6. Conclusões e discussões

Os agrupamentos gerados pelo K-means apresentaram bons resultados, com a escolha otimizada de k resultando em clusters coerentes, conforme a matriz de distâncias e a dispersão dos pontos. A segmentação refletiu padrões identificados na análise dos dados, evidenciando a robustez do modelo.

Entre os 10 países mais felizes estão Dinamarca, Suíça, Islândia, Noruega, Finlândia, Canadá, Países Baixos, Nova Zelândia, Austrália e Suécia. A única mudança no período analisado ocorreu em 2019, quando a Áustria substituiu a Austrália.

Já os países menos felizes apresentam maior variação e instabilidade. A maioria está na África Subsaariana, com algumas exceções no Oriente Médio, Norte da África, América Latina e Sul da Ásia. Ao longo dos anos, Chade, Guiné, Afeganistão, Síria, Burundi e Togo figuraram frequentemente entre os menos felizes, enquanto Madagascar, Tanzânia, Haiti, Iémen, Botswana, Malawi e Zimbabwe entraram no ranking em diferentes momentos. Isso sugere que PIB per capita, saúde, suporte social e mudanças socioeconômicas são fatores críticos para a segmentação.

Além disso, os países mais felizes apresentam pouca variação no score, refletindo maior estabilidade, enquanto os menos felizes exibem flutuações significativas, conforme ilustrado a seguir:

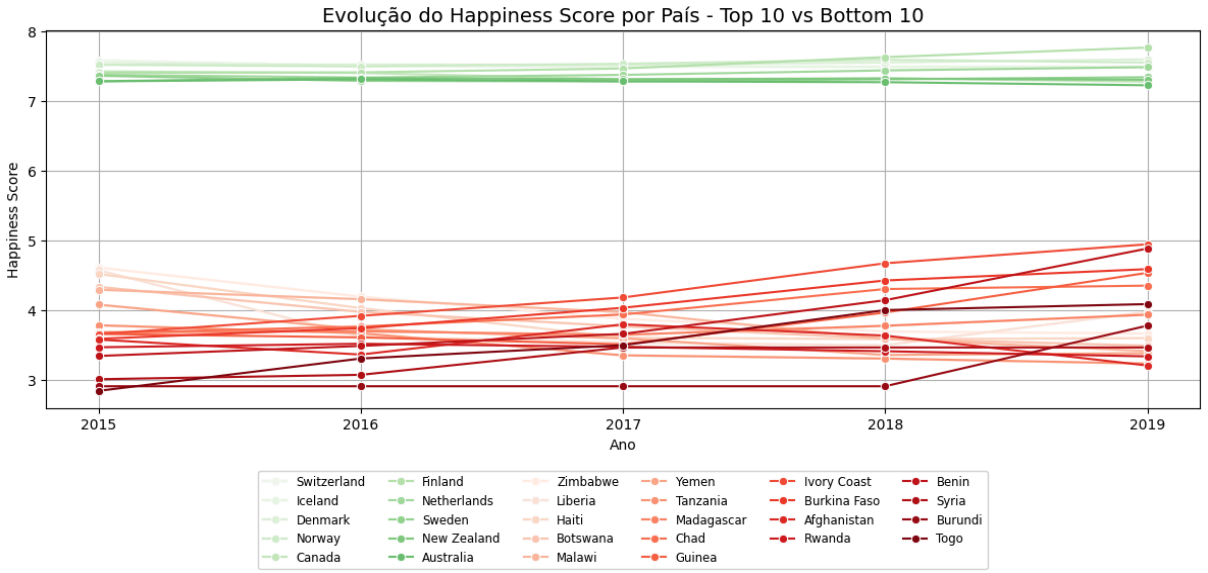

Figura 7: gráfico da evolução do score dos 10 países mais felizes e menos felizes (2015-2019)

No caso do agrupamento geral com a média do período, os resultados refletiram com consistência os padrões anuais, mostrando um agrupamento similar ao encontrado a cada ano e mantendo as distribuições regionais nos *clusters*.

Dessa forma, esses achados reforçam a forte influência do PIB *per capita*, suporte social e expectativa de vida na felicidade das populações, onde as regiões desenvolvidas (Europa Ocidental, América do Norte e Oceania) estão predominantemente no *cluster* de maior felicidade. Os países em desenvolvimento são agrupados em um cluster intermediário, dada a ampla margem dos indicadores de felicidades. Isso resulta em um cluster diverso com várias regiões e países. Os países menos desenvolvidos, principalmente na África Subsaariana e partes da Ásia e Oriente Médio, compõem o *cluster* de menor felicidade.

7. Referências

Fontana, A. e Naldi, M. C. (2009) "Estudo de Comparação de Métodos para Estimação de Números de Grupos em Problemas de Agrupamento de Dados". Universidade de São Paulo. ISSN - 0103-2569.

IBM. (2024) "O que é agrupamento K-means?". Disponível em: https://www.ibm.com/br-pt/topics/k-means-clustering. Acesso em: 12 mar. 2025.

KAGGLE. (2024) "World Happiness Report dataset". Disponível em: https://www.kaggle.com/datasets/unsdsn/world-happiness/. Acesso em: 12 mar. 2025.

Vassalo, D. H. C. (2023) "Análises de competições presentes na Plataforma Kaggle para auxiliar no desenvolvimento de novas soluções para problemas de visão computacional." Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) - Instituto de Computação, Universidade Federal de Alagoas, Maceió, 2021.

World Happiness Report. (2024) "About the World Happiness Report." Disponível em: https://worldhappiness.report/about/. Acesso em: 12 mar. 2025.

8. Trabalhos relacionados

POLIS, Kaori Tobara Felippe. “Comparação de métodos de análise de cluster para agrupar países de acordo com os dados do Relatório Mundial da Felicidade.” Orientadora: Miriam Rodrigues Silvestre. 2023. 58 f. Trabalho de Conclusão de Curso (Bacharelado em Estatística) - Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista, Presidente Prudente, 2023.

BIZARRIA, Fabiana Pinto de Almeida; BARBOSA, Flávia Lorenne Sampaio; ROCHA, Soraia Germana de Sousa. “Considerações sobre Felicidade por Clusters de discentes de administração em uma Instituição Pública de Ensino Superior.” *Revista Brasileira de Administração Científica*, v. 8, n. 1, 2017. Disponível em: https://www.sustenere.inf.br/index.php/rbadm/article/view/SPC2179-684X.2017.001.0009. Acesso em: 17 mar. 2025.

DETRINIDAD, Emmanuel; LÓPEZ-RUIZ, Víctor-Raúl. “The interplay of happiness and sustainability: a multidimensional scaling and K-means cluster approach.” *Sustainability*, v. 16, n. 22, p. 10068, 2024. Disponível em: https://www.mdpi.com/2071-1050/16/22/10068. Acesso em: 17 mar. 2025.