**Aprendizado de Máquina 1**

**Trabalho 2**

Aprendizado não Supervisionado - “Unsupervised Learning on Country Data”

**Professor:** Prof. Dr. Murilo Naldi

**Grupo:**

Augusto dos Santos Gomes Vaz - 800268

Gabriel Lourenço de Paula Graton - 800432

Pietro Minghini Moralles - 792238

Vitor Matheus da Silva - 800260

São Carlos

22/08/2023

**Índice**

[**1. Resumo 2**](#_lasf0skr71zt)

[**2. Introdução 2**](#_t0v0edy1v3cs)

[**3. Escolha do Conjunto de Dados 3**](#_87jrq0tq9adg)

[3.1. Justificativa da escolha do conjunto de dados 3](#_y93jti7yq61f)

[**4. Estudo do Conjunto de Dados 4**](#_oya1cp13rgiz)

[4.1. Entendendo o conjunto de dados 4](#_waopq7ti1icw)

[4.2. Imputação de novos objetos 5](#_dzt0nelc10ye)

[4.3. Análise descritiva dos dados 6](#_i5ujlnlqc51)

[4.4. Tratamento de Outliers 14](#_uquhrl8d912g)

[4.5. Normalização 14](#_63iezotk6fbq)

[4.6. Correlação entre os atributos 15](#_4207kvymsvfq)

[4.7. Redução de Dimensionalidade 17](#_ydt3a63j1ksg)

[**5. Agrupamento 19**](#_dtiyfxggt19m)

[5.1. Justificativa da escolha dos métodos de agrupamento 19](#_icad9zpwocub)

[5.2. Aplicação dos Algoritmos 19](#_90q1s6w2fyly)

[**6. Validação 25**](#_2hh34asye060)

[**7. Mapas 27**](#_t7jbzabw7val)

[**8. Conclusão 31**](#_ft2sxfvvgg4h)

[**9. Bibliografia 33**](#_pygc6dje2yf0)

# Resumo

Inicialmente, o grupo conduziu uma pesquisa com o objetivo de determinar o conjunto de dados mais adequado e relevante para o projeto. Esse processo levou em consideração não apenas as informações contidas, mas também a maneira como a análise exploratória de dados seria conduzida, as etapas de pré-processamento necessárias e a finalidade do conjunto de dados em questão.

Após essa etapa, era de suma importância considerar como a exploração seria conduzida em diferentes etapas, visando uma seleção mais criteriosa de atributos para obter classes mais coerentes.

Por fim, foi decidido adotar abordagens de tratamento de dados, normalização e a utilização de algoritmos não supervisionados, incluindo a elaboração de gráficos de dispersão que representassem as estratégias empregadas. Além disso, realizou-se uma comparação de desempenho entre os cenários apresentados. O processo foi concluído com a representação das informações contidas no conjunto de dados por meio de amostragem em mapas.

# Introdução

Esse trabalho foi feito com objetivo de colocar em prática os ensinamentos passados em sala de aula na disciplina Aprendizado de Máquina I pelo professor Murilo Coelho Naldi, referente aos conteúdos de Aprendizado não supervisionado. Serão utilizados algoritmos em um conjunto escolhido pelo grupo, a fim de permitir o agrupamento, obtendo informações de gráficos de visualizações e medidas de validação, que serão discutidas adiante.

Dessa forma, propomos neste trabalho um estudo acerca do dataset “Unsupervised Learning on Country Data”, para que seja possível utilizar as estratégias aprendidas em aula, sem um classificação já estabelecida. Além disso, é preciso entender o que realmente é relevante para o problema apresentado, como visualizar os dados e verificar a presença de outliers e como realmente tratar as informações e proceder a partir disso.

# Escolha do Conjunto de Dados

Para a realização do projeto, o grupo optou por selecionar o conjunto de dados “Unsupervised Learning on Country Data”, disponível no repositório de dados Kaggle. O conjunto de dados pode ser acessado por meio do seguinte link: [Link para o Conjunto de Dados](https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data).

O desafio proposto com base nesse conjunto de dados envolve uma situação na qual a empresa “HELP International” arrecadou aproximadamente $10 milhões de dólares com o objetivo de prestar assistência a países em necessidade. Nesse contexto, a tarefa consiste em categorizar os países usando indicadores de saúde e fatores socioeconômicos, a fim de avaliar o nível de desenvolvimento de cada nação. Esse processo de categorização visa possibilitar o uso estratégico e eficaz dos investimentos obtidos.

O conjunto de dados contém 167 objetos, sendo cada um destes países, os quais contém 10 atributos cada que serão explicados e discutidos posteriormente. Todas as manipulações dos dados e aplicações podem ser encontradas no notebook do Collab, que pode ser acessado neste link: [Link da realização do trabalho](https://colab.research.google.com/drive/15870fTQmsyFDvg5lwFegQeiIW0UtpJQV?usp=sharing)

## Justificativa da escolha do conjunto de dados

Optamos por utilizar esse conjunto de dados principalmente devido à complexidade das relações entre suas variáveis, uma vez que fatores socioeconômicos e de saúde estão intrinsecamente interligados. Além disso, a tarefa de categorização nesse conjunto de dados possui implicações concretas no mundo real, uma vez que o modelo não supervisionado pode auxiliar na identificação de grupos de países que compartilham características semelhantes, proporcionando uma alocação mais eficaz de recursos.

Adicionalmente, a abordagem não supervisionada tem o potencial de revelar padrões e agrupamentos que não seriam aparentes de outra forma, o que poderia conduzir descobertas inesperadas e insights que contribuem para o conhecimento geral sobre o desenvolvimento global.

Quando nos deparamos com esse dataset, fomos instigados por suas complexidade e particularidades. Esse conjunto de dados permite a aplicação da linguagem Python, juntamente com suas diversas bibliotecas, ao mesmo tempo que possibilita a aplicação de conceitos que foram ensinados em aula. Resumidamente, escolhemos o dataset pelos seguintes motivos:

* Identificação de atributos relevantes para agrupamento;
* Ampliação dos dados para uma visão abrangente;
* Compreensão dos dados;
* Tratamento de outliers;
* Relevância da classificação no contexto global.

# Estudo do Conjunto de Dados

### Entendendo o conjunto de dados

Conforme mencionado anteriormente, o conjunto de dados abrange 167 entradas, cada uma composta por um total de 10 atributos. Esses atributos consistem em indicadores de saúde e fatores socioeconômicos. É crucial destacar que os dados foram coletados, mas não há indício do ano que foi feito a coleta - há uma variância entre 2005 e 2010 , o que pode resultar em disparidades em relação ao cenário atual. Os atributos estão listados a seguir:

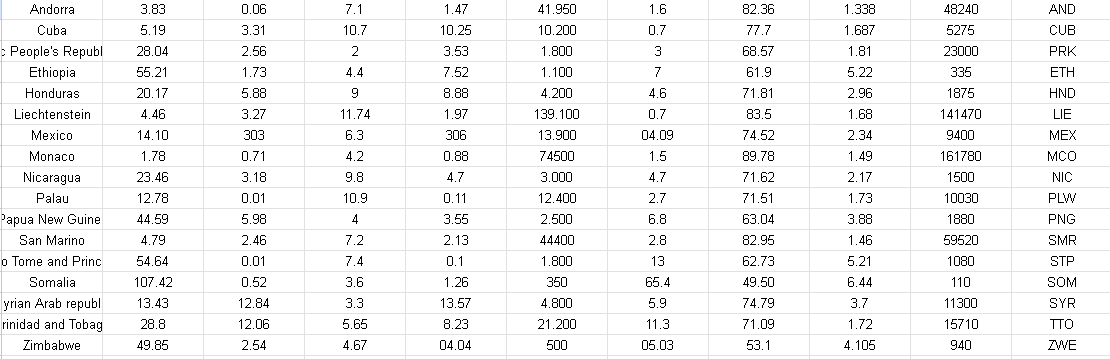
* **country:** representa o país do qual os dados foram coletados.
* **child\_mort:** representa o número de mortes de crianças com menos de 5 anos a cada mil nascimentos.
* **exports:** representa a exportação de bens e serviços. Expresso como porcentagem do PIB per capita.
* **health:** representa os gastos totais com saúde por habitante. Expresso como porcentagem do PIB per capita
* **imports:** representa a importação de bens e serviços por habitante. Expresso como porcentagem do PIB per capita.
* **income:** representa a renda líquida por pessoa.
* **inflation:** representa a medição do aumento contínuo dos preços de bens e serviços na economia.
* **life\_expect:** representa a média de anos que uma criança recém-nascida viveria se os padrões de mortalidade atuais se mantivessem os mesmos.
* **total\_fer:** representa o número médio de crianças que nasceriam para cada mulher se as taxas atuais de idade e fertilidade permanecessem as mesmas.
* **gdpp:** representa o PIB per capita. Calculado como o PIB Total dividido pela população total.

Além dos atributos originais da base de dados, também foi incluído o atributo "code", que consiste em códigos geográficos curtos desenvolvidos para representar países e áreas dependentes. A necessidade desse atributo surgiu para facilitar a representação dos clusters em um mapa, visando tornar a diferença entre os países mais interativa e perceptível.

### Imputação de novos objetos

Uma das primeiras observações feitas pelo grupo foi a quantidade de entradas no conjunto de dados. Neste conjunto, há um total de 167 países representados. No entanto, em 2023, existem aproximadamente 193-196 países reconhecidos internacionalmente pela ONU. Diante dessa constatação, decidimos realizar pesquisas para coletar informações sobre os países ausentes.

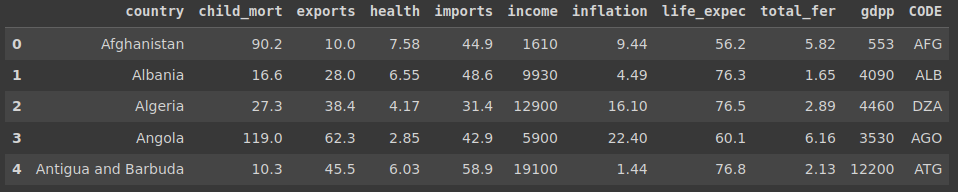
Os dados coletados pelo grupo referem-se ao ano de 2010, uma vez que ao compará-los com os dados presentes no conjunto, foram encontradas semelhanças. No entanto, devido à falta de especificação da fonte original, a coleta de informações para os países ausentes na base de dados tornou-se uma tarefa desafiadora. Após a obtenção dos novos dados, o grupo chegou à conclusão de que não seria viável incorporar os novos países ao conjunto de dados. Isso se deve ao fato de que as discrepâncias entre as fontes estavam resultando em valores altamente inconsistentes, o que poderia potencialmente enviesar os algoritmos a serem aplicados no futuro.



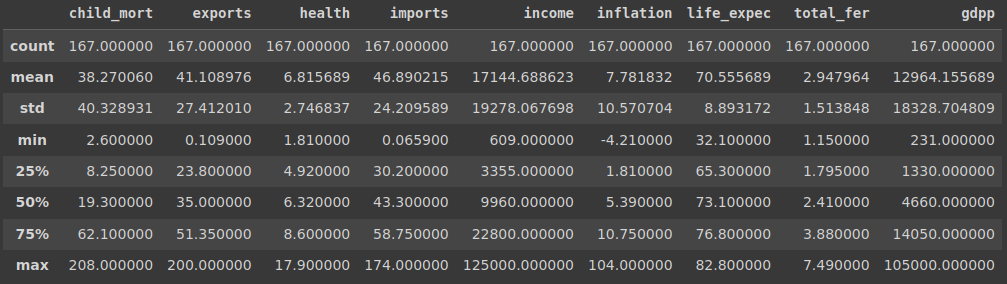
***Figura 1:*** *Tentativa de imputação de dados*

### Análise descritiva dos dados

Nas duas figuras abaixo, é possível observar um breve resumo de como os dados estão distribuídos na base de dados, e uma descrição dos valores numéricos presentes. Na figura que descreve os valores numéricos, é possível observar que esses valores destoam bastante entre si. Portanto, se faz necessária a realização da normalização, que será comentada mais adiante.

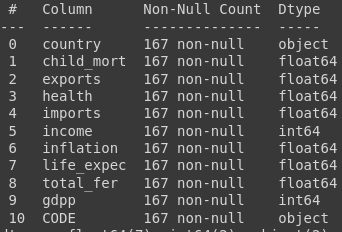


***Figura 2:*** *Head do conjunto de dados*



***Figura 3:*** *Descrição dos valores numéricos*

Também foram realizadas as checagens para encontrar possíveis dados duplicados ou faltantes, os quais retornaram negativos para ambos os casos. A figura abaixo demonstra a natureza dos dados contidos no conjunto, sendo a maioria desses valores numéricos, enquanto apenas “country” e “CODE” representam valores categóricos, os quais serão removidos para a aplicação dos algoritmos de agrupamento, pois servem apenas como auxiliares para a utilização em gráficos.



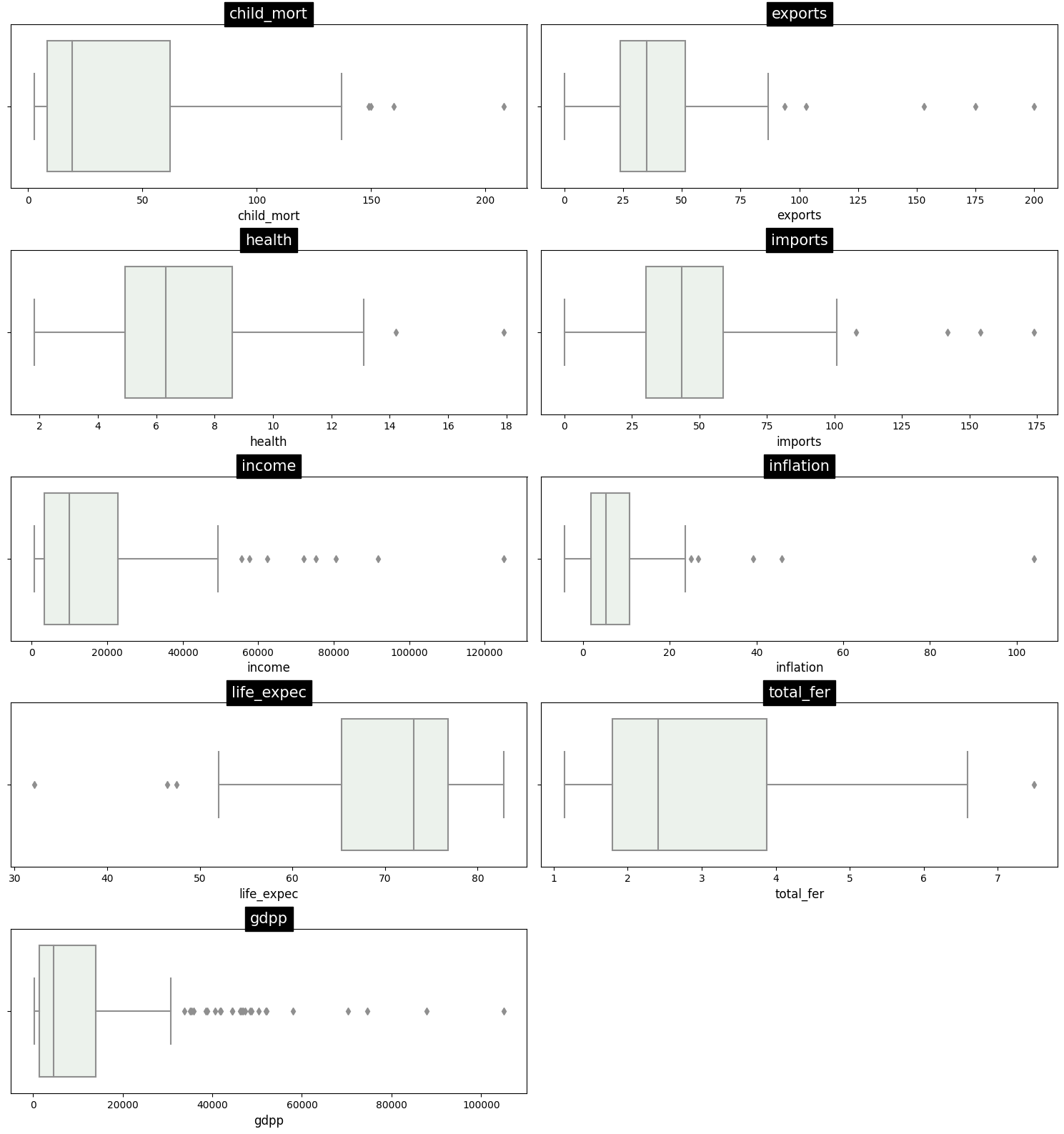
***Figura 4:*** *Natureza dos dados*

Nesta figura, apresentamos um box plot que descreve a distribuição dos valores numéricos presentes na base de dados. O box plot é uma representação gráfica que exibe a mediana, quartis e possíveis valores discrepantes. Observamos que os valores numéricos exibem uma grande variação entre si, o que indica a necessidade de normalização para melhor comparabilidade entre os países.

Vale ressaltar que nos atributos como exportação, importação, renda e PIB per capita, há a presença de valores bastante discrepantes, indicando a existência de outliers no banco de dados. No nosso caso, a figura sugere que alguns países possuem valores muito acima da média, o que pode afetar as análises estatísticas.

No atributo "expectativa de vida", uma discrepância interessante merece destaque, na qual os valores isolados estão à esquerda do diagrama de caixas. Isso sugere que um grupo de países apresenta expectativas de vida significativamente mais baixas em comparação com a maioria dos outros países. A disposição dos valores no diagrama de caixas auxilia na identificação visual dessas diferenças.

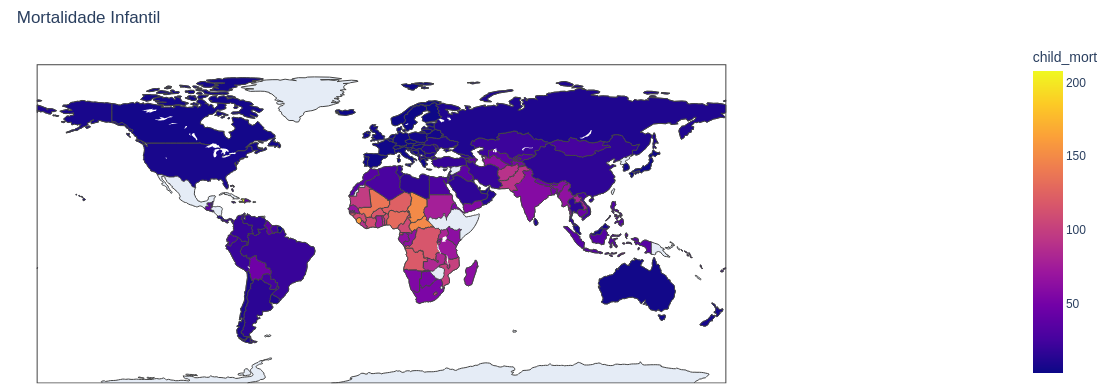
Essas observações obtidas a partir dos diagramas de caixas enfatizam a necessidade de uma análise mais aprofundada para compreender as razões subjacentes a essas discrepâncias e outliers, e como eles podem influenciar as conclusões derivadas dos dados.



***Figura 5:*** *BoxPlot dos atributos do conjunto de dados*

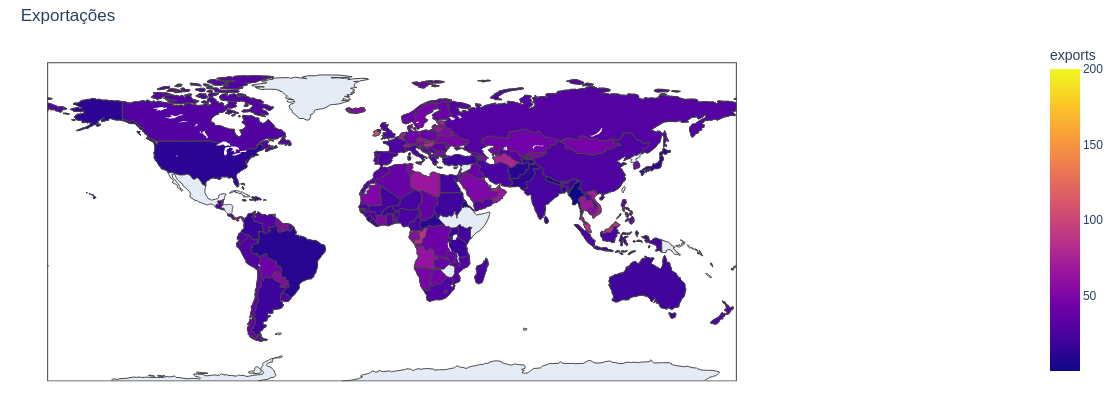
Após a análise da distribuição estatística dos atributos, foram gerados gráficos para observar, de maneira interativa e mais visual, como os mesmos estão distribuídos. Vale ressaltar que foram utilizados apenas os atributos numéricos.

* Atributo ‘**child\_mort**’’: o gŕafico abaixo representa o número de mortes de crianças com menos de 5 anos a cada mil nascimentos.



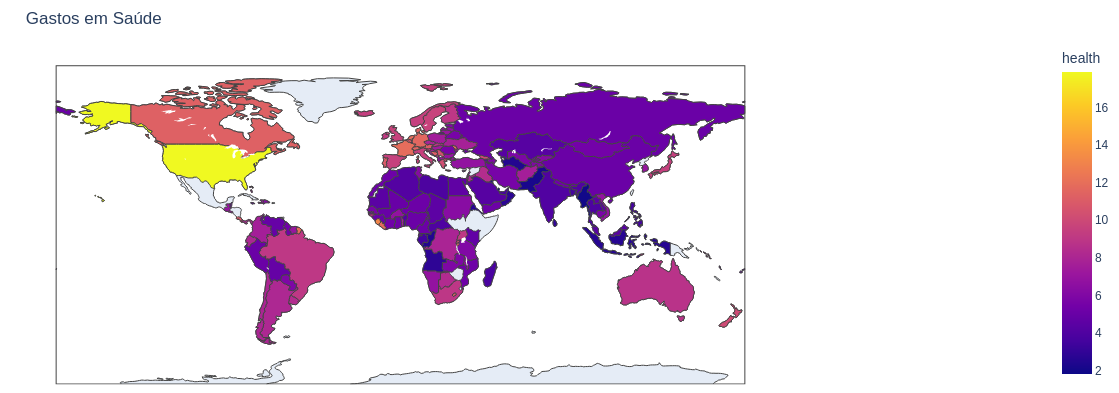
***Figura 6 :*** *mortalidade infantil a cada mil*

* Atributo ‘**exports**’: o gráfico abaixo representa a exportação de bens e serviços. Expresso como porcentagem do PIB per capita.

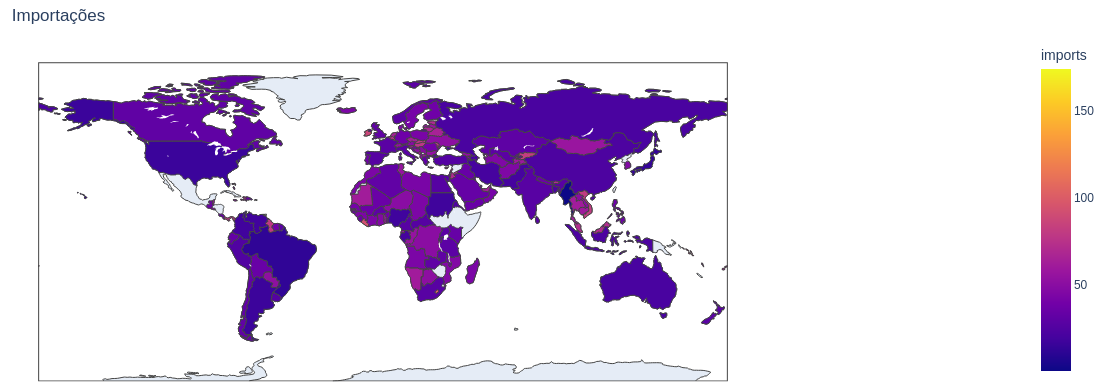
****

***Figura 7:*** *Exportações em % PIB per capita*

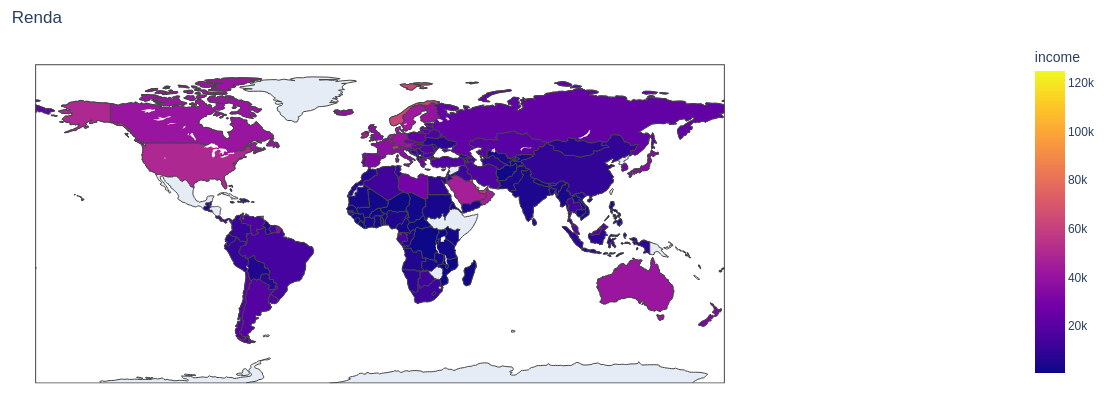
* Atributo ‘**health**’: o gráfico abaixo representa os investimentos saúde por habitante. Expresso como porcentagem do PIB per capita

***Figura 8:*** *Investimento em saúde em % PIB per capita*

* Atributo ‘**imports**’: o gráfico abaixo representa a importação de bens e serviços por habitante. Expresso como porcentagem do PIB per capita.

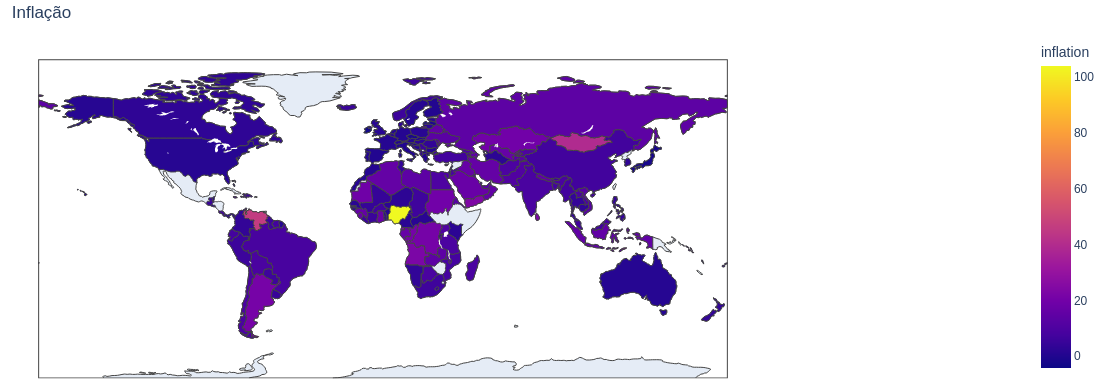
***Figura 9:*** *Importações em % PIB per capita*

* Atributo ‘**income**’: o gráfico abaixo representa a renda líquida por pessoa.



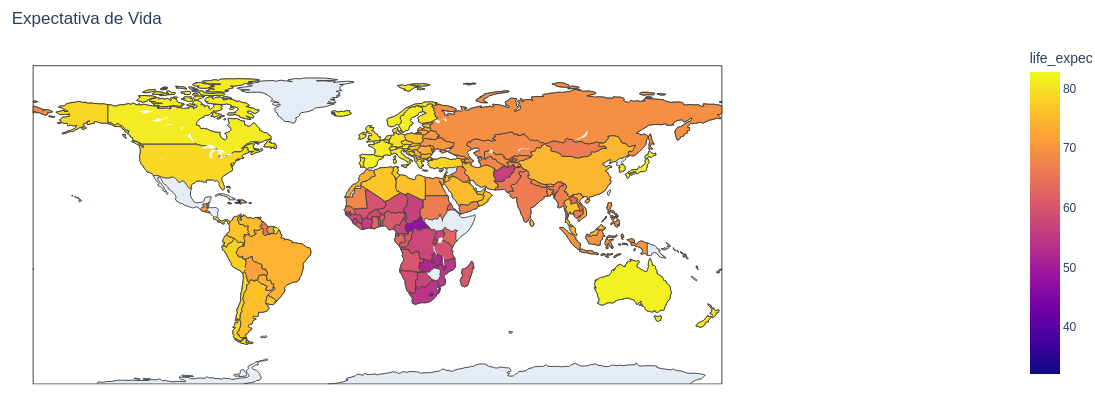
***Figura 10:*** *Renda per capita anual em dólares*

* Atributo ‘**inflation**’: o gŕafico abaixo representa a medição do aumento contínuo dos preços de bens e serviços na economia.

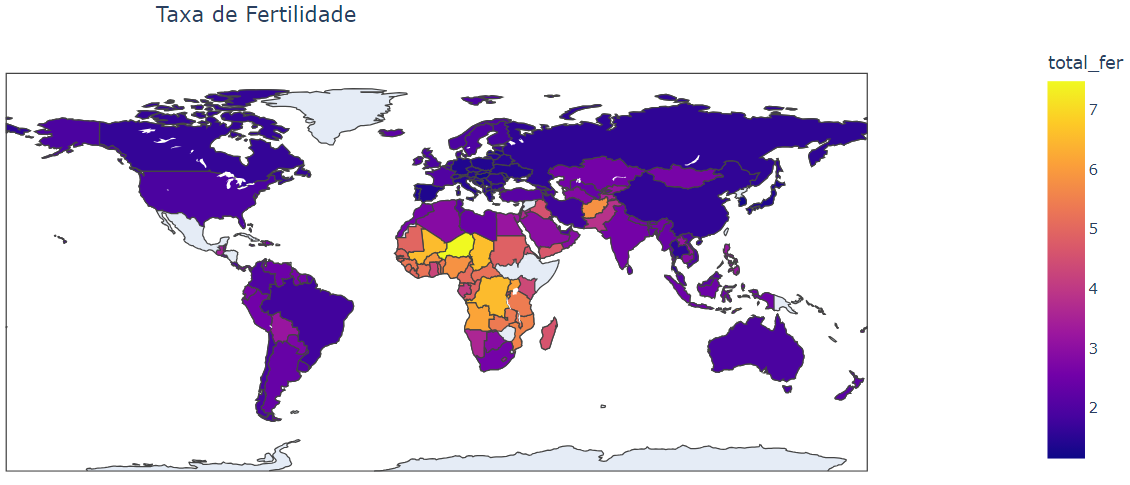


***Figura 11:*** *inflação em %*

* Atributo ‘**life\_expect**’: o gráfico abaixo representa a média de anos que uma criança recém-nascida viveria se os padrões de mortalidade atuais se mantivessem os mesmos.

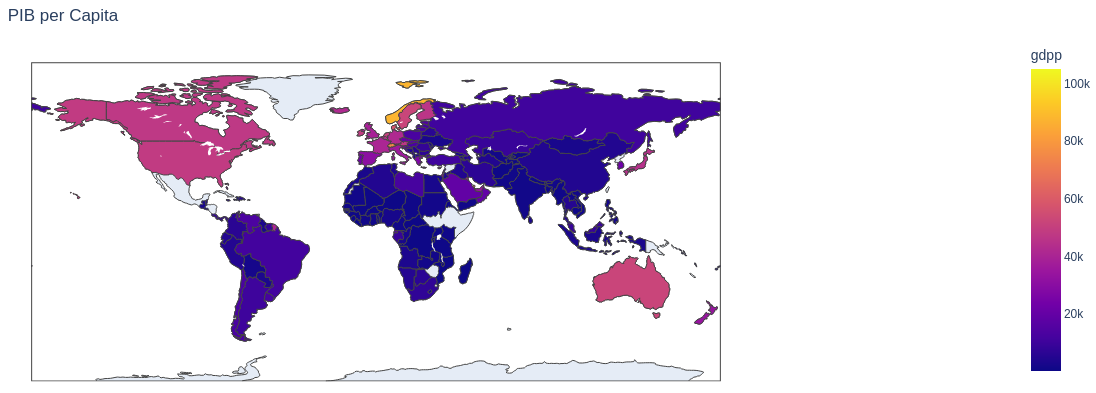
***Figura 12:*** *Expectativa de vida em anos*

* Atributo ‘**total\_fer**’: o gráfico abaixo representa o número médio de crianças que nasceriam para cada mulher se as taxas atuais de idade e fertilidade permanecem as mesmas.



***Figura 13:*** *Taxa de fertilidade em nascidos por mil*

* Atributo ‘**gdpp**’: o gráfico abaixo representa o PIB per capita. Calculado como o PIB Total dividido pela população total.

****

***Figura 14:*** *PIB per capita em dólares*

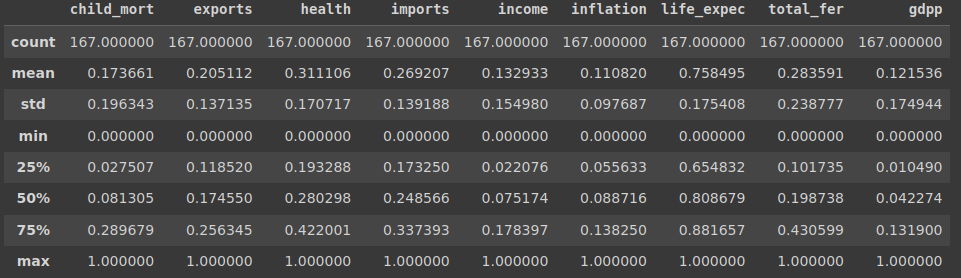
### Tratamento de Outliers

Como mencionado anteriormente, no conjunto de dados estão presentes alguns países cujos valores diferem significativamente dos demais. Embora possam ser considerados outliers, que possivelmente prejudicam o desempenho dos algoritmos de agrupamento, optamos por mantê-los. Essa decisão foi baseada no princípio de que, mesmo que esses valores se apresentem muito distintos, eles ainda refletem a realidade socioeconômica do mundo. Em outras palavras, esses valores são válidos e correspondentes a situações reais.

### Normalização

Realizamos a normalização dos atributos por meio do MinMaxScaler, uma vez que esses atributos apresentam escalas muito distintas. Com esse procedimento, ajustamos os valores de diferentes atributos para uma escala que varia de 0 a 1, de modo a evitar distorções na análise devido às diferenças de magnitude entre os atributos. Além disso, essa normalização se mostrou necessária, uma vez que os algoritmos de clusterização empregados fundamentam-se na similaridade (distância) euclentre os objetos. Dessa maneira, a normalização foi adotada para impedir que aqueles atributos com escalas maiores exerçam um impacto desproporcional nas medidas utilizadas nos algoritmos.

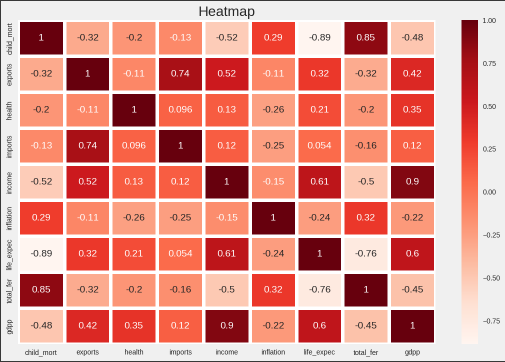
Na figura abaixo, podemos observar como os atributos ficaram em uma escala bem mais reduzida após a aplicação da normalização.



***Figura 15:*** *Conjunto de dados após a normalização*

### Correlação entre os atributos

Após a normalização, submetemos os atributos a um heatmap no qual vamos analisar a correlação entre eles. A correlação é uma medida estatística utilizada para avaliar a relação entre duas variáveis, indicando a força e a direção da associação entre essas variáveis. Na figura a seguir, apresentamos a correlação entre os atributos do conjunto de dados.



***Figura 16:*** *Matriz de correlação entre os atributos*

Diante desta imagem, podemos inferir as seguintes conclusões:

* Os pares de variáveis que mais se correlacionam positivamente são child\_mort e total\_fer (0.85), exports e imports (0.74), exports e income (0.52), income e gdpp (0.9), income e life\_expec (0.61), life\_expec e gdpp (0.66).
* Os pares de variáveis que mais se correlacionam negativamente são child\_mort e life\_expec, income e child\_mort, life\_expec e total\_fer.
* De uma maneira geral, os atributos possuem forte correlação entre si, ao passo que, na vida real, isso também é válido. É possível citar como exemplo a correlação entre mortalidade infantil e expectativa de vida. Quanto maior a mortalidade infantil em um país, menor será a expectativa de vida, além de outros atributos que também podem ser citados

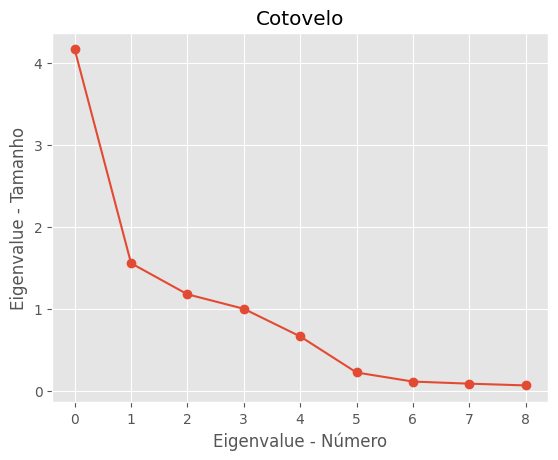
### Redução de Dimensionalidade

Desenvolvendo o HeatMap e conduzindo a análise, torna-se evidente a presença de diversas correlações marcantes entre os atributos, destacadas pela intensidade das cores representadas. Alguns exemplos incluem a correlação entre a taxa de mortalidade infantil e a expectativa de vida, bem como entre a taxa de fecundidade e a renda líquida. Além disso, são observadas relações entre exportações e importações, assim como entre expectativa de vida e taxa de fecundidade.

Devido a essa situação, algumas variáveis podem conter informações semelhantes, resultando em redundância e ineficácia no modelo. Dessa forma, a multicolinearidade se manifesta entre os atributos. Para mitigar esse problema, optou-se pela aplicação da Análise de Componentes Principais (PCA). Essa abordagem revela-se vantajosa nesse contexto, pois reduz a dimensionalidade ao criar novas variáveis chamadas componentes principais. Esses componentes são combinações lineares dos atributos originais, contribuindo para eliminar a redundância e a multicolinearidade, ao mesmo tempo que preserva a maior parte da variação original dos dados.

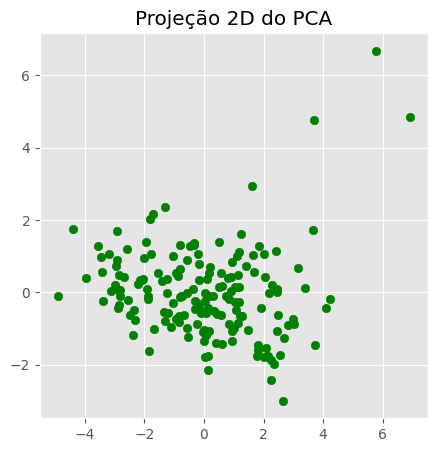
Seguindo essa lógica, a escolha recaiu sobre a configuração padrão de inicialização do PCA, porém com a definição de k=2. Essa decisão baseia-se em diversos motivos:

Optamos por selecionar os componentes até o ponto de inflexão identificado no Gráfico Abaixo. No referido gráfico, essa inflexão é claramente observada no índice 2. Consequentemente, nossa escolha recaiu sobre os componentes nos índices 0, 1 e 2, totalizando assim dois componentes no conjunto selecionado.



***Figura 17:*** *Determinação do número de dimensões*

Além disso, ao adotar essa quantidade específica de componentes, torna-se viável visualizar os elementos que serão aglutinados posteriormente em um gráfico de dispersão. Esse gráfico permite a identificação clara dos valores discrepantes presentes, bem como dos agrupamentos (clusters) que têm potencial para se formar.

****

***Figura 18:*** *Projeção do Dataset com dimensões reduzidas*

# Agrupamento

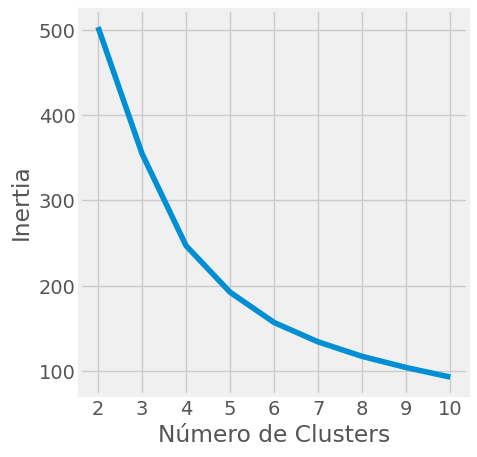
## Justificativa da escolha dos métodos de agrupamento

Após a análise exploratória dos dados e todas as transformações necessárias, os algoritmos de agrupamento foram aplicados. O grupo optou por utilizar três algoritmos: K-Means, Algoritmo Hierárquico Aglomerativo Ward e Algoritmo Hierárquico Aglomerativo Complete. A escolha por esses algoritmos se deve ao fato de terem sido ensinados em sala de aula, e desejamos praticar a aplicação deles.

## Aplicação dos Algoritmos

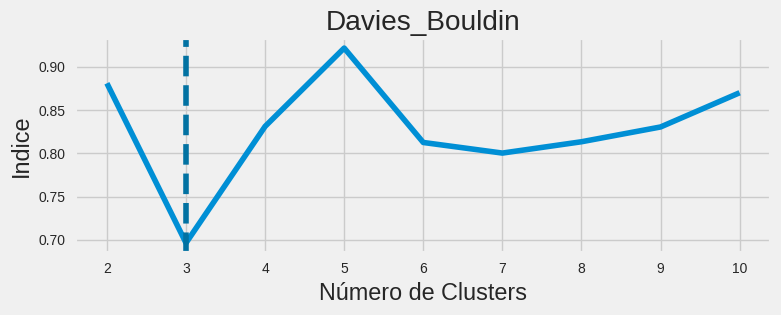
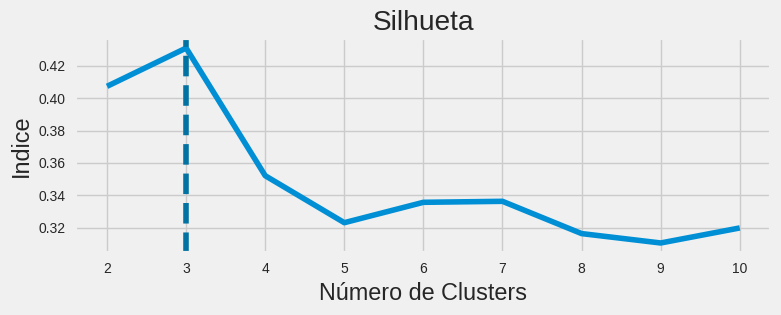
**a. Kmeans**

Inicialmente, com o objetivo de determinar o número ideal de clusters para o conjunto de dados normalizado e reduzido em suas dimensões, executamos o algoritmo variando de 2 a 11 clusters. Isso nos permitiu criar um gráfico que relaciona a variabilidade intra-cluster com o número de clusters aplicados. O gráfico abaixo ilustra essa análise no conjunto de dados, indicando que o número ideal de clusters é 5.



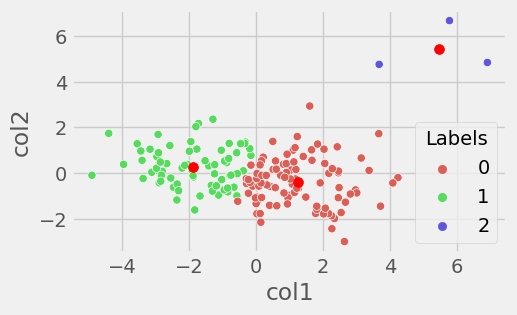
***Figura 19:*** *Método do Cotovelo - KMeans*

Além disso, optamos por utilizar as métricas de Silhueta e Davies Bouldin para confirmar o número ideal de clusters. A primeira métrica informa a semelhança entre os objetos dentro de um cluster e a dissimilaridade em relação a outros clusters, enquanto a segunda leva em consideração a média da distância entre os centros dos clusters e a dissimilaridade entre os pontos de dados dentro desses clusters. Os gráficos a seguir demonstram que, em ambas as métricas, o número ideal de clusters é considerado 3.

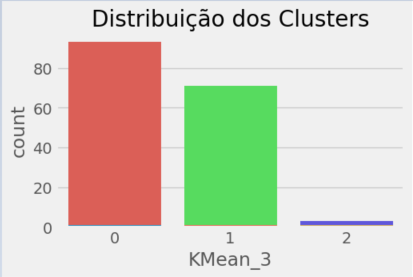


***Figura 20:*** *Silhueta e Davies-Bouldin - KMeans*

Diante dessa situação, a equipe optou por prosseguir com a implementação do algoritmo KMeans utilizando 5 e 3 clusters, uma vez que as métricas revelaram discrepâncias notáveis entre essas configurações. As primeiras visualizações consistem na representação gráfica da distribuição dos pontos, seguida pelo diagrama de dispersão correspondente. Nessas representações, é perceptível que certas nações foram alocadas em grupos exclusivos, identificados como outliers. Por outro lado, as outras duas colunas no diagrama apresentam semelhanças notáveis.

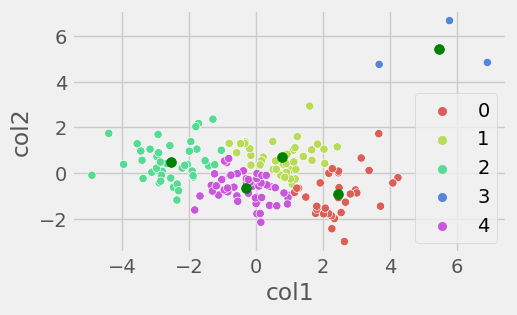


***Figura 21:*** *Diagrama de Pontos - KMeans3*

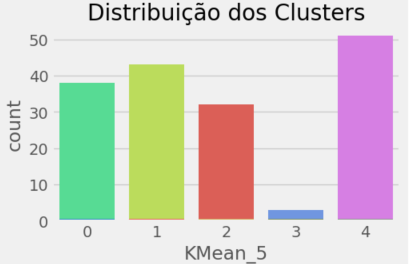


***Figura 22:*** *Distribuição dos pontos - KMeans3*

Conforme mencionado anteriormente, também foi adotada a abordagem de 5 clusters. Embora os mesmos países peculiares identificados anteriormente estejam presentes no mesmo grupo, a distribuição dos demais pontos resultou em quatro divisões mais distintas. Nesse caso, observa-se uma maior homogeneidade entre as mesmas



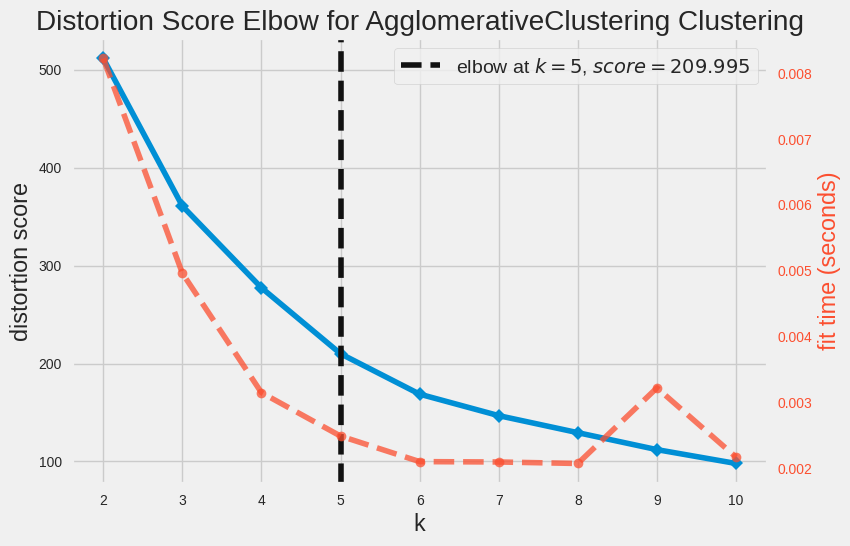
***Figura 23:*** *Diagrama de Pontos - KMeans5*



***Figura 24:*** *Distribuição dos pontos - KMeans5*

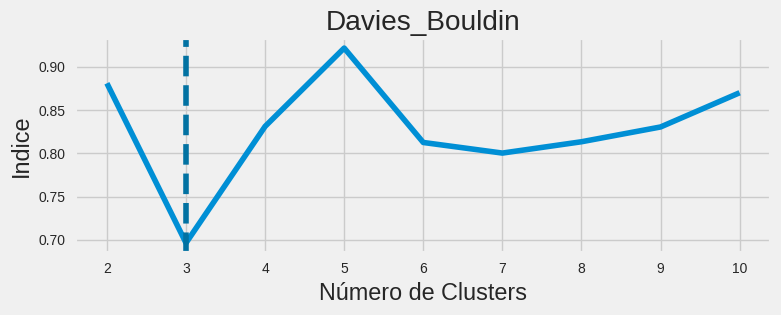
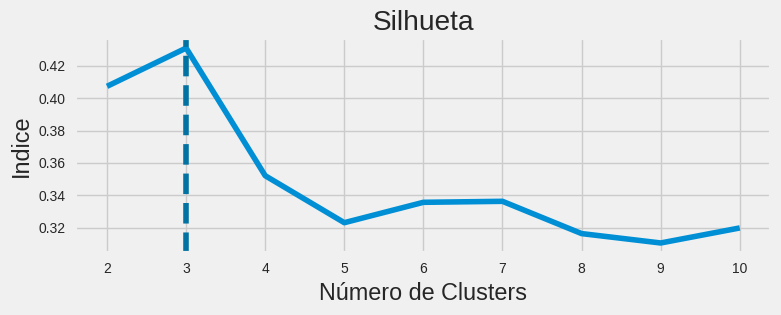
**b. Aglomerativo**

Realizamos uma abordagem idêntica no que diz respeito à quantidade ideal de clusters para a utilização dos algoritmos aglomerativos. Inicialmente, o algoritmo foi executado com um número de clusters variando de 2 até 11, a fim de identificarmos o número ideal de clusters usando o método do cotovelo. O gráfico a seguir demonstra que o ponto de cotovelo está identificado no número de clusters igual a 5.



***Figura 25:*** *Método do Cotovelo - Aglomerativo*

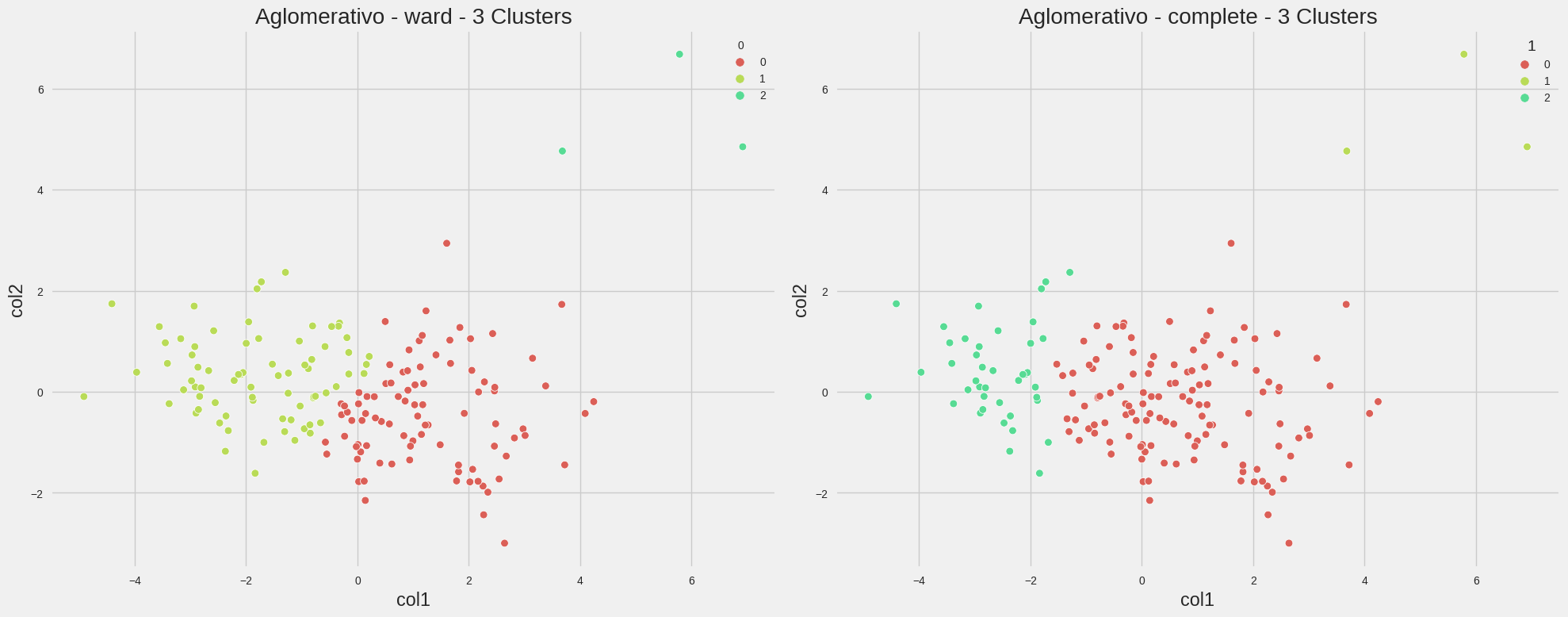
Além disso, também foram realizadas as análises das métricas de Silhueta e Davies Bouldin. Os gráficos abaixo demonstram que o número ideal de clusters é 3 em ambas as medidas.



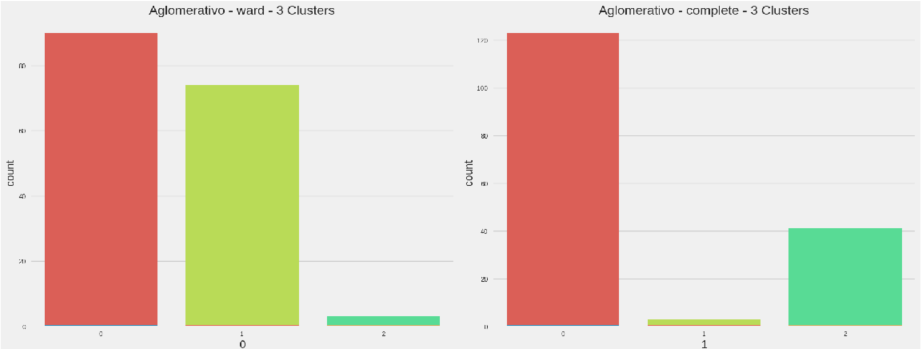
***Figura 26:*** *Silhueta e Davies-Bouldin -Aglomerativo*

Portanto, a equipe optou por aplicar o algoritmo de agrupamento aglomerativo com 5 e 3 clusters, com o objetivo de observar as diferenças entre as abordagens. Dentre as diversas estratégias possíveis de agrupamento aglomerativo, foram selecionadas apenas as técnicas "ward" e "complete". Essa seleção se justifica pelo fato de que somente essas duas estratégias são suficientes para ilustrar como diferentes abordagens dentro do mesmo método podem resultar em diferenças visíveis.

Inicialmente, a abordagem com 3 clusters foi executada. Essa configuração distribui os pontos de forma semelhante à estratégia anterior. Dois grupos apresentam certa semelhança entre si, enquanto um terceiro grupo contém os outliers, representados pelos três primeiros países. Esses aspectos são claramente evidenciados nas duas imagens a seguir.

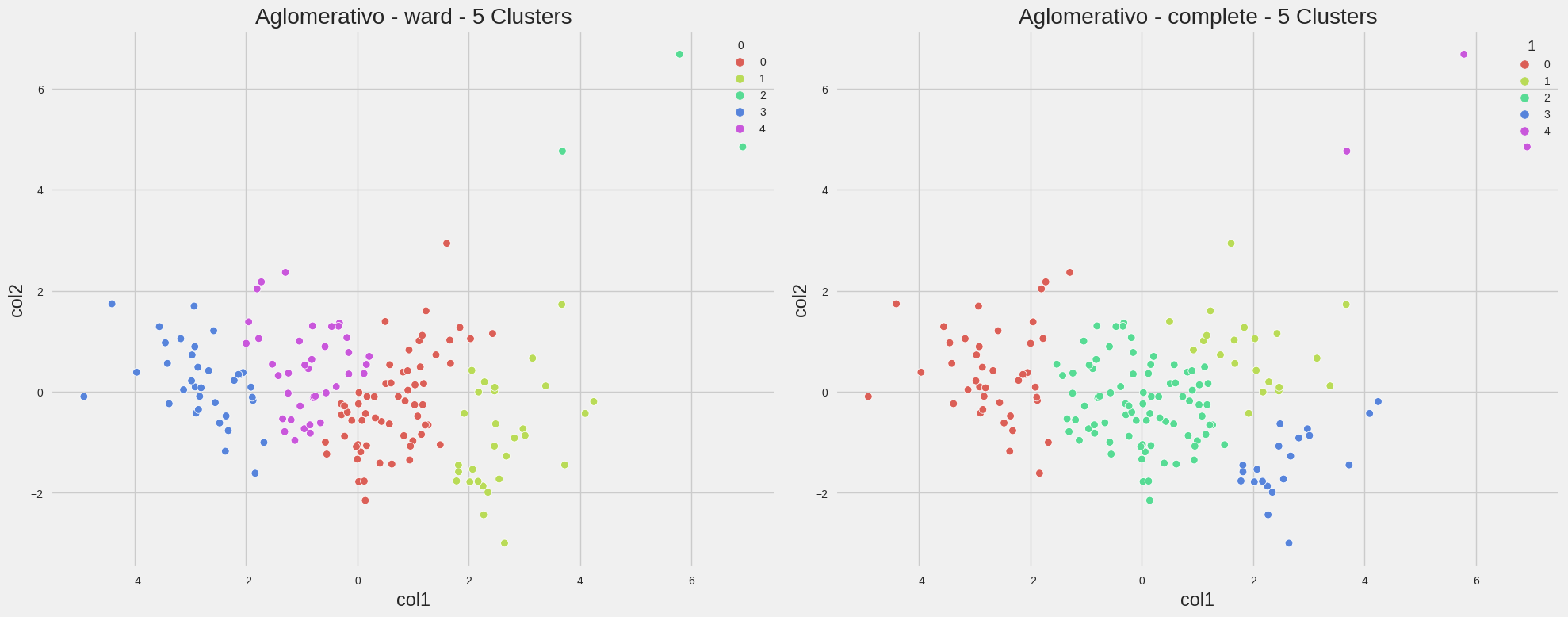
****

***Figura 27:*** *Diagrama de Pontos - Agg3*

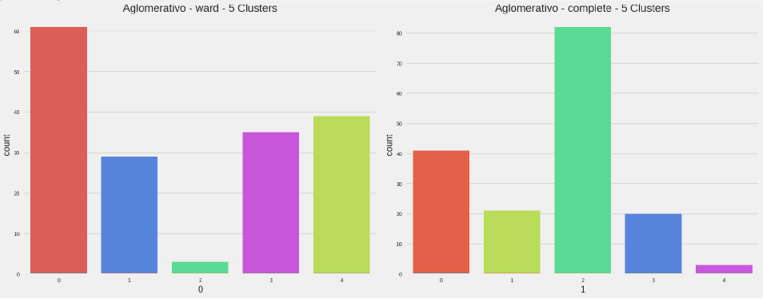
****

***Figura 28:*** *Distribuição dos pontos - Agg5*

Posteriormente, o processo aglomerativo foi executado para 5 clusters. De maneira geral, há semelhanças com as abordagens anteriores, além da presença de um grupo adicional. É possível notar que, apesar dessa diferenciação, uma linearidade entre os clusters ainda é mantida, embora haja sempre um grupo discrepante que se destaca em relação aos demais.

****

***Figura 29:*** *Diagrama de Pontos - Agg5*

****

***Figura 30:*** *Distribuição dos pontos - Agg5*

# Validação

Como dito anteriormente, a validação rápida foi feita antes de escolher cada cluster. Entre as validações escolhidas estão a métrica Silhueta. Por essa métrica, quanto menor a distância do objeto para o seu grupo e maior a distância do objeto para os outros grupos , melhor.

Mas também a métrica de Davies-Bouldin que mede a dispersão interna de um cluster e a dispersão entre os clusters vizinhos. Quanto menor o valor do índice, mais coesos e bem separados estão os clusters.

| **Métrica de Validação: Silhueta** | | | |
| --- | --- | --- | --- |
| **Agrupamento** | **Algoritmo** | **Silhueta** | **Davies-B** |
| Particional | K-Means (3) | 0.44250 | 0.68242 |
| Particional | K-Means (5) | 0.38206 | 0.79265 |
| Hierárquico | Aglomerativo completo (3) | 0.43387 | 0.62331 |
| Hierárquico | Aglomerativo ward (3) | 0.43084 | 0.69631 |
| Hierárquico | Aglomerativo completo (5) | 0.38639 | 0.73399 |
| Hierárquico | Aglomerativo ward (5) | 0.32315 | 0.92161 |

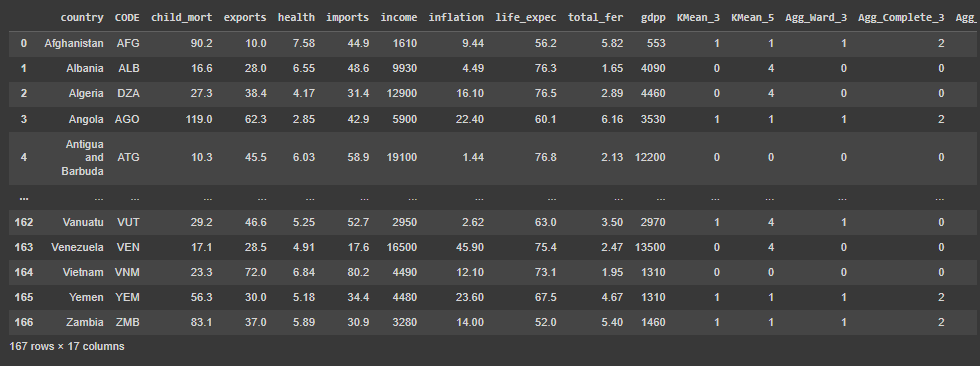
Dessa maneira, ao examinarmos os dados através das avaliações realizadas, as estratégias que foram adotadas visando obter as métricas mais favoráveis demonstraram um desempenho superior nesse contexto. No entanto, quando comparamos as duas abordagens distintas de agrupamento, as métricas de silhueta e Davies-Bouldin sobressaíram na estratégia de KMeans com 3 clusters.

Adicionalmente, é importante notar que o índice Davies-Bouldin apresentou um aumento expressivo na abordagem "ward" com 5 clusters. Este índice teve uma queda abrupta ao passar de 3 para 5 clusters na eficiência , revelando-se como o pior valor em ambas as abordagens.

# Mapas

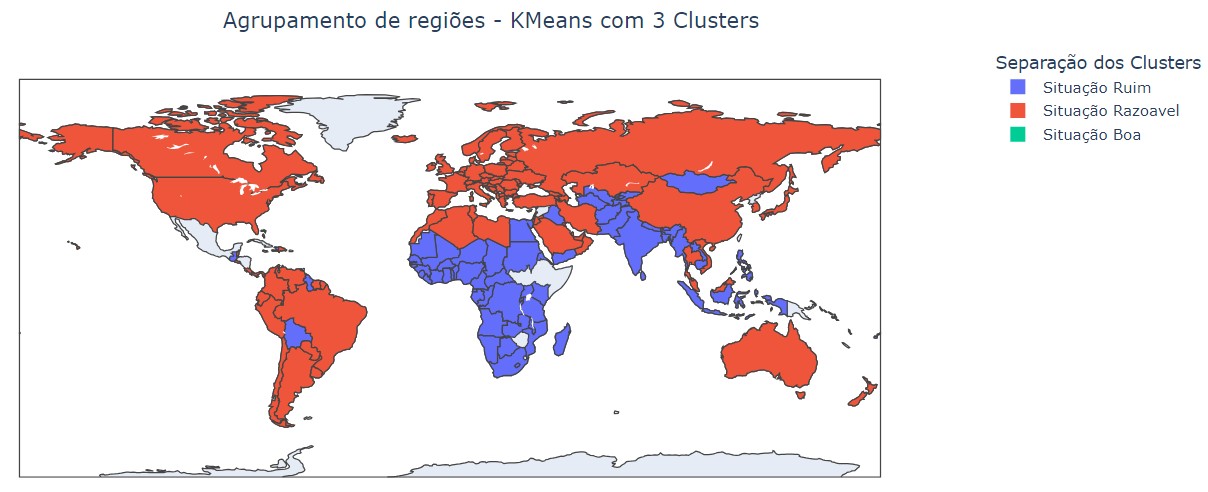
Com a aplicação dos algoritmos e suas respectivas validações realizadas, o grupo optou por realizar a separação dos clusters em um mapa, de modo a deixar mais vísivel o agrupamento realizado em cada país. Para isso, fizemos a utilização da API Chrolopleth Map, utilizada para representar a variação espacial baseada em quantidades.

A identificação de cada país no mapa foi feita com os clusters obtidos na aplicação dos algoritmos, juntamente com a coluna CODE, utilizada para indexar cada mapa em um cluster específico. O conjunto de dados final pode ser observado na figura abaixo.

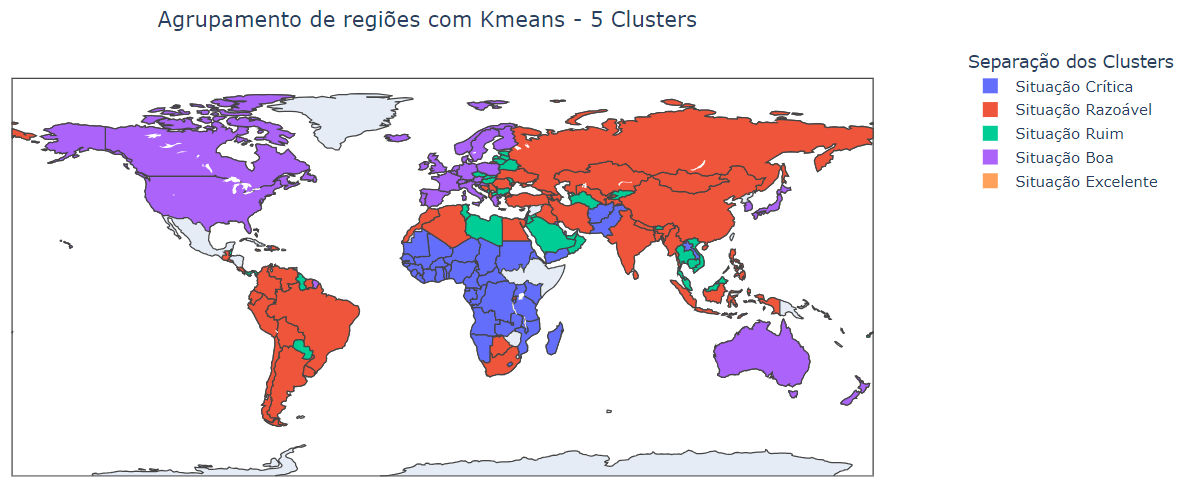


***Figura 31:*** *Dataset Final*

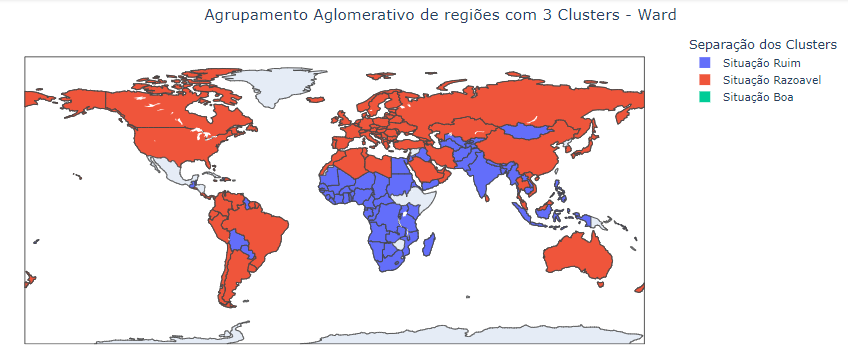
Feito isso, segue abaixo as imagens respectivas de cada mapa respectivamente identificado pelo algoritmo aplicado.



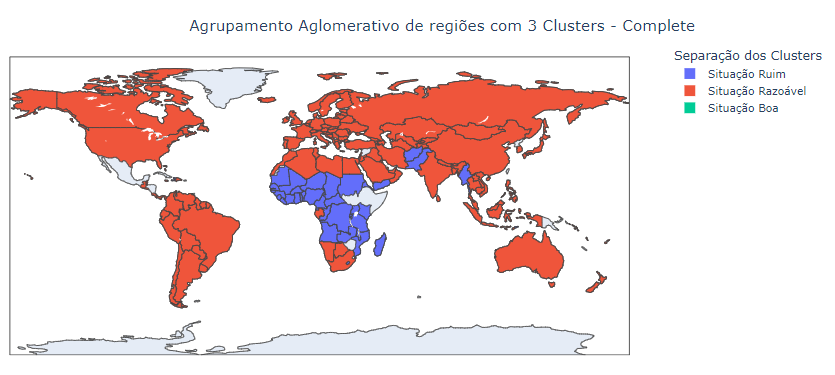
***Figura 32:*** *Mapa das regiões - KMeans3*



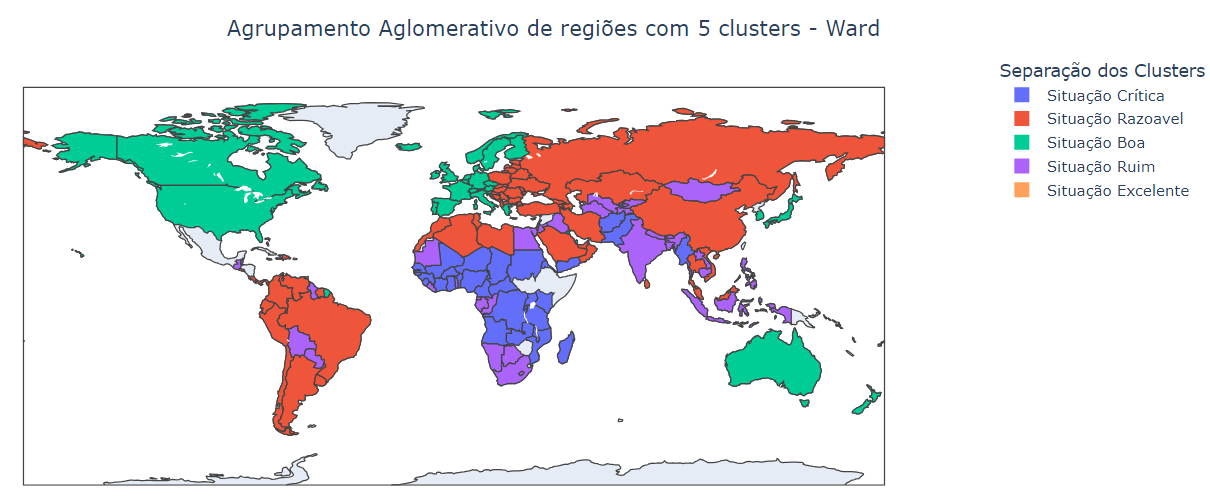
***Figura 33:*** *Mapa das regiões - KMeans5*



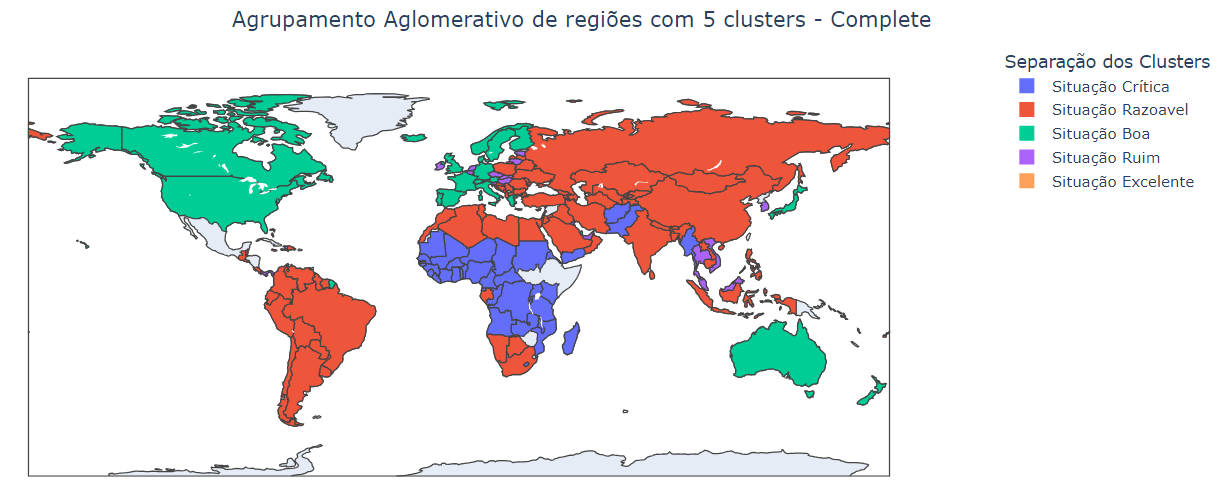
***Figura 34:*** *Mapa das regiões - Agg Ward 3*



***Figura 35:*** *Mapa das regiões - Agg Complete 3*



***Figura 36:*** *Mapa das regiões - Agg Ward 5*



***Figura 37:*** *Mapa das regiões - Agg Complete 5*

Com estes mapas, é possível observar alguns padrões quanto ao agrupamento dos países. De certa maneira, os países da América do Norte, Europa Ocidental, Austrália e Nova Zelândia, em todos os agrupamentos, foram classificados como países em situações socioeconômicas boas, o que indica que estes não precisam de investimentos ou ajudas da organização HELP International. Em contrapartida, os países da África Subsariana foram classificados como os mais necessitados de ajuda na maioria dos algoritmos.

# Conclusão

Dado o exposto, é possível verificar os processos que foram necessários para trabalhar com o conjunto de dados escolhido, o "Country Data". O conjunto de dados passou por uma pesquisa inicial para entender do que se trata e quais dados ele contém. Inicialmente, foi considerada a ideia de imputação de dados, seguida pela análise dos outliers e de como abordá-los, seja mantendo-os ou excluindo-os. Tudo isso foi feito para compreender quais países precisam de assistência em um contexto socioeconômico.

Após essas etapas, foram escolhidos os melhores tipos de gráficos para representar cada atributo. Esses gráficos permitem visualizar os intervalos das taxas de atributos para diferentes países. Além disso, foram utilizados gráficos de box plot para identificar rapidamente os outliers visíveis. Em seguida, um mapa de calor foi empregado para mostrar a correlação entre os atributos, a qual se mostrou significativa.

Com todas essas informações, aplicou-se a Análise de Componentes Principais (PCA) para auxiliar na visualização dos dados e para mitigar o erro proveniente da alta correlação entre os atributos. Posteriormente, foram aplicados os algoritmos K-Means, respeitando as validações e também considerando o método do cotovelo. Da mesma forma, os algoritmos aglomerativos foram aplicados respeitando o método do cotovelo e as validações correspondentes.

Além disso, as comparações das validações realizadas resultaram na criação de gráficos baseados nos clusters. Com isso, foi possível fazer comparações com os Índices de Desenvolvimento Humano (IDHs) e mapas de desenvolvimento, buscando validar os agrupamentos realizados.

Por fim, vale ressaltar que, dependendo do número de clusters aplicados, é possível perceber que o agrupamento deixa de ser efetivo e começa a unir objetos de maneira inadequada, fazendo com que países em situações distintas sejam agrupados juntos. Por outro lado, se o número de clusters for muito pequeno, ocorre pouca distinção entre os países, o que dificulta a identificação das nações que necessitam de investimentos.

# Bibliografia

* 1. <https://data.worldbank.org/indicator>
  2. <https://www.indexmundi.com>
  3. <https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data>
  4. <https://drive.google.com/drive/folders/1vWlEDfs6uKf_Nc-lVqajW-aSsMYg8lel>