



CESAR SCHOOL
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Probabilidade e Estatística

Turma A - 2024.1

II Avaliação

Relatório Técnico - Data Set

Alunos:

Ana Beatriz Ximenes Alves

Caio Barreto de Albuquerque

RECIFE

2024

SUMÁRIO

INTRODUÇÃO.....	3
1. IMPORTANDO FERRAMENTAS NECESSÁRIAS.....	5
2. COMO RODAR EM SUA MÁQUINA.....	7
3. ANÁLISES INICIAIS.....	8
3.1. Estatística Descritiva.....	8
A. Histograma Inicial.....	8
B. Box Plot.....	9
C. Histogramas específicos (por coluna).....	10
D. Mapa de Calor.....	11
E. Scatter Plot.....	12
3.2. Normalização.....	16
A. Normalização (Parte Matemática).....	16
B. Dados Matemáticos.....	17
3.3. Distribuição Amostral.....	17
A. Line Plot (Comparação entre países).....	20
3.4. Intervalo de Confiança.....	21
4. ANÁLISES APROFUNDADAS.....	25

INTRODUÇÃO

O dataset escolhido contém informações sobre o Índice de Desenvolvimento Humano (IDH) de vários países ao longo de diferentes anos, desde 1990 até 2022, fornecendo, portanto, uma visão abrangente da evolução do IDH e de outros indicadores relacionados à qualidade de vida e desenvolvimento humano.

Nele, temos as seguintes informações, em formato de colunas:

- IDH por ano: Dados do IDH de cada país em diversos anos (1990, 2000, 2010, 2015, 2019, 2020, 2021, 2022).
- Crescimento médio do IDH: Taxas de crescimento médio do IDH entre períodos específicos (2000-2010, 2010-2022, 1990-2022).
- Indicadores de desenvolvimento humano adicionais: Expectativa de vida ao nascer, anos esperados de escolaridade, anos médios de escolaridade, renda nacional bruta per capita e outros indicadores relevantes.

Com a análise da evolução do IDH e seus componentes ao longo do tempo para identificar padrões de desenvolvimento, desigualdades entre países, e o impacto de diferentes fatores no desenvolvimento humano, podemos entender melhor como políticas e eventos históricos influenciaram a qualidade de vida globalmente. Para realizar as análises, utilizaremos dos indicadores disponibilizados por meio de:

- Análise Descritiva: Para resumir as principais características dos dados e fornecer uma visão geral da distribuição dos indicadores.
- Visualização de Dados: Gráficos de linhas e barras para mostrar a evolução temporal do IDH e comparações entre países.
- Análise de Correlação: Para identificar relações entre diferentes indicadores, como IDH e renda per capita.
- Distribuições: Avaliar a dispersão e centralização dos dados de IDH e outros indicadores.

Ao final do trabalho, pretendemos concluir com base em Tendências Temporais, identificando se houve melhorias ou retrocessos no desenvolvimento humano ao longo das décadas, realizando comparações entre os países, destacando quais tiveram os maiores e menores crescimentos no IDH e explorar possíveis razões para essas variações, e, por fim, os impactos de Políticas e Eventos, discutindo como diferentes políticas, crises econômicas ou eventos globais podem ter afetado o desenvolvimento humano.

Em suma, este relatório fornecerá uma análise detalhada do desenvolvimento humano global, destacando áreas de progresso e necessidade de melhorias, contribuindo para a formulação de políticas e estratégias de desenvolvimento mais eficazes. O dataset analisado está disponível no link

https://drive.google.com/file/d/1GZDGzNfNKIRz3SE3IrMk3XPieZyqbSZ-/view?usp=share_link e o google colab no qual foram feitas as codificações que auxiliaram a análise no link https://colab.research.google.com/drive/1OnzOur8a4GKBBhhZtn10oWNAxM_7wQT2?usp=sharing.

1. IMPORTANDO FERRAMENTAS NECESSÁRIAS

Para a análise de dados realizada, foi essencial importar e utilizar várias bibliotecas do ecossistema Python, cada uma com funcionalidades específicas que contribuíram para diferentes aspectos do trabalho. Abaixo, detalhamos as bibliotecas utilizadas, suas finalidades, e as funções específicas que foram relevantes para nossa análise.

A. Pandas

```
import pandas as pd
```

Pandas é uma biblioteca de software escrita para a linguagem de programação Python, usada para manipulação e análise de dados, oferecendo estruturas de dados e operações para manipular tabelas numéricas e séries temporais. Ela é amplamente utilizada para limpeza de dados, preparação de dados, análise exploratória de dados (EDA) e operações complexas de dados, como junções, filtragens e agrupamentos. Sendo assim, Pandas foi crucial para carregar, limpar e manipular os dados do nosso dataset, facilitando a realização de operações complexas de dados de forma eficiente e intuitiva.

B. Numpy

```
import numpy as np
```

Numpy é uma biblioteca fundamental para computação científica em Python, fornecendo suporte para arrays e matrizes multidimensionais, além de uma coleção de funções matemáticas para operar esses arrays. É usada principalmente para realizar cálculos numéricos de forma eficiente, sendo a base de muitas outras bibliotecas de ciência de dados e aprendizado de máquina em Python. Sendo assim, Numpy foi utilizada para cálculos matemáticos e estatísticos, como a média e o desvio padrão, necessários para a análise de dados e para calcular intervalos de confiança.

C. Matplotlib

```
import matplotlib.pyplot as plt
```

Matplotlib é uma biblioteca de plotagem em 2D do Python, que produz figuras de alta qualidade em uma variedade de formatos gráficos, usada para criar gráficos estáticos, animados e interativos em Python. Por ser altamente customizável e permitindo a criação de

visualizações complexas, foi usada no projeto para criar visualizações dos intervalos de confiança das variáveis numéricas, permitindo a plotagem de distribuições normais e a visualização dos intervalos de confiança de forma clara e informativa.

D. Seaborn

```
import seaborn as sns
```

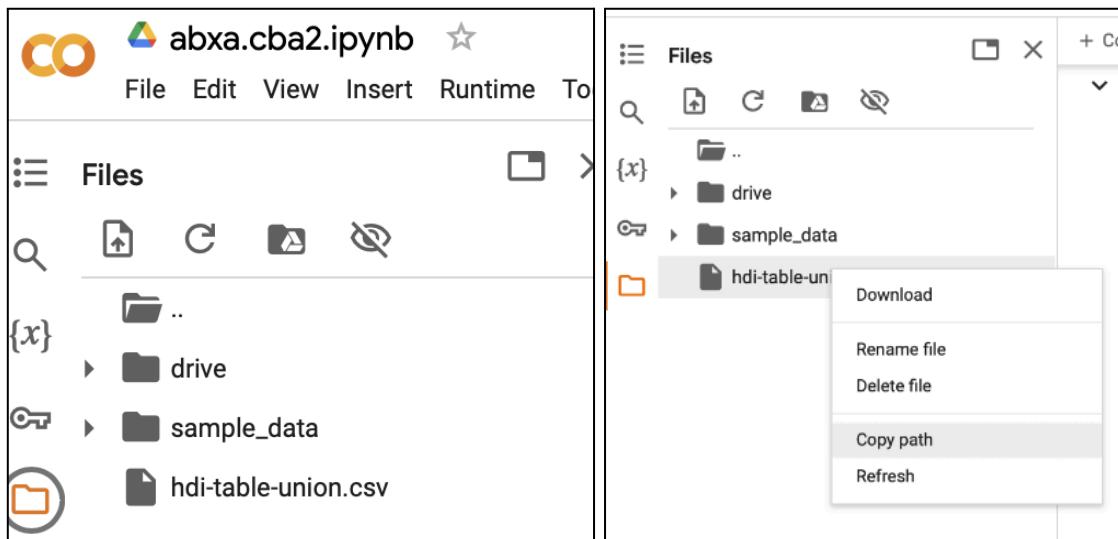
Seaborn é uma biblioteca de visualização de dados baseada no Matplotlib que fornece uma interface de alto nível para desenhar gráficos estatísticos atraentes e informativos, usada para facilitar a criação de gráficos complexos e informativos. Por se integrar-se bem com Pandas, permitindo criar visualizações com menos código comparado ao Matplotlib, foi utilizado para criar gráficos de distribuição que destacaram os intervalos de confiança das variáveis, simplificando a visualização de dados estatísticos complexos.

Sendo assim, cada uma dessas ferramentas desempenhou um papel crucial no processamento, análise e visualização dos dados. Pandas e Numpy foram essenciais para a manipulação e cálculo dos dados, enquanto Matplotlib e Seaborn facilitaram a criação de visualizações informativas. O Google Drive, embora não explicitamente usado no código, forneceu um meio eficaz de armazenar e compartilhar recursos do projeto, promovendo a colaboração eficiente.

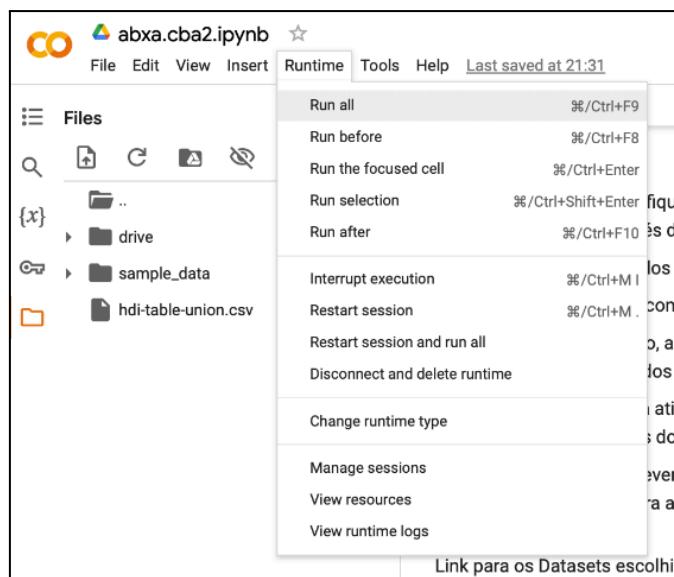
2. COMO RODAR EM SUA MÁQUINA

Para replicar a análise de dados descrita ao longo do trabalho, o primeiro passo é obter o dataset que será utilizado, que foi disponibilizado via link do Google Drive, que permite o acesso direto ao arquivo, seja para visualização ou download.

Após o download, importe para o ambiente de análise de dados, conforme demonstrado abaixo:



Após adicionar no ambiente de desenvolvimento, aperte na opção de “Run All”, para que todos os gráficos e cálculos sejam realizados automaticamente.

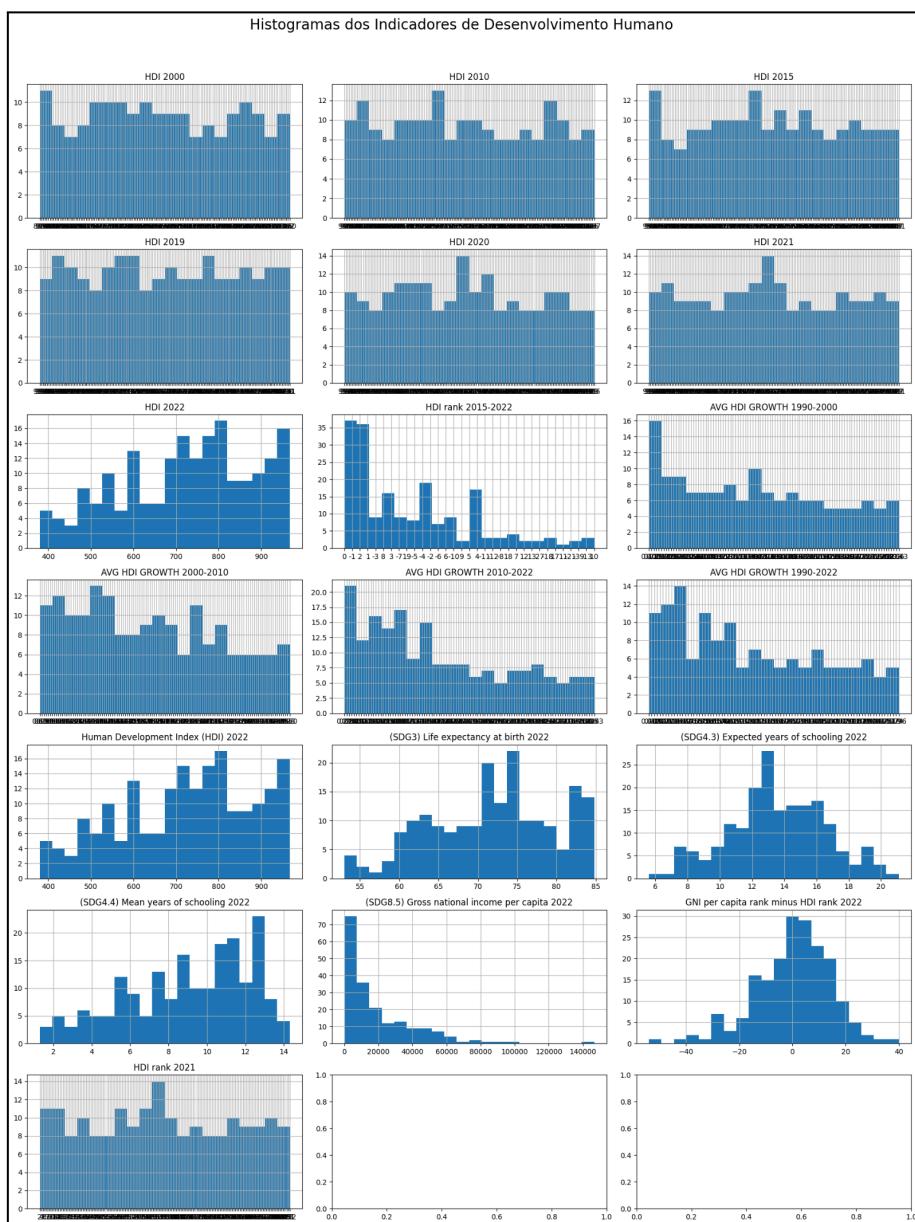


3. ANÁLISES INICIAIS

Conforme os requisitos solicitados para a entrega, e baseando-se nos exemplos apresentado em aulas, realizamos algumas explorações dos dados, que tiveram como intuito gerar inferências fáticas sobre o dataset acima apresentado e descrito. Abaixo, destrinchamos os resultados das explorações feitas.

3.1. Estatística Descritiva

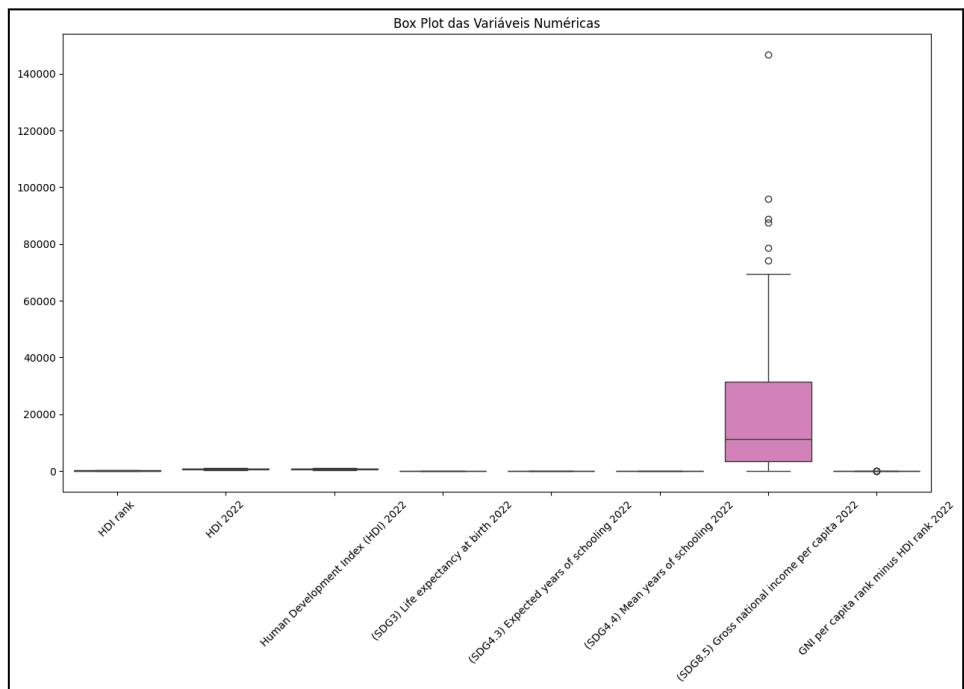
A. Histograma Inicial



Importante relembrar que um histograma é um tipo de gráfico utilizado para representar a distribuição de um conjunto de dados, agrupando os dados em intervalos (ou bins) e mostrando a frequência de valores em cada intervalo. Dessa forma, no contexto de análise de dados, os histogramas são extremamente úteis para visualizar a distribuição de variáveis e identificar padrões.

Vejamos que na primeira tentativa de aplicação, ocorreu que, em razão dos intervalos utilizados serem muito pequenos (em razão da alta quantidade de países agrupados), o resultado foi uma visualização que dificultou a interpretação da distribuição geral dos dados.

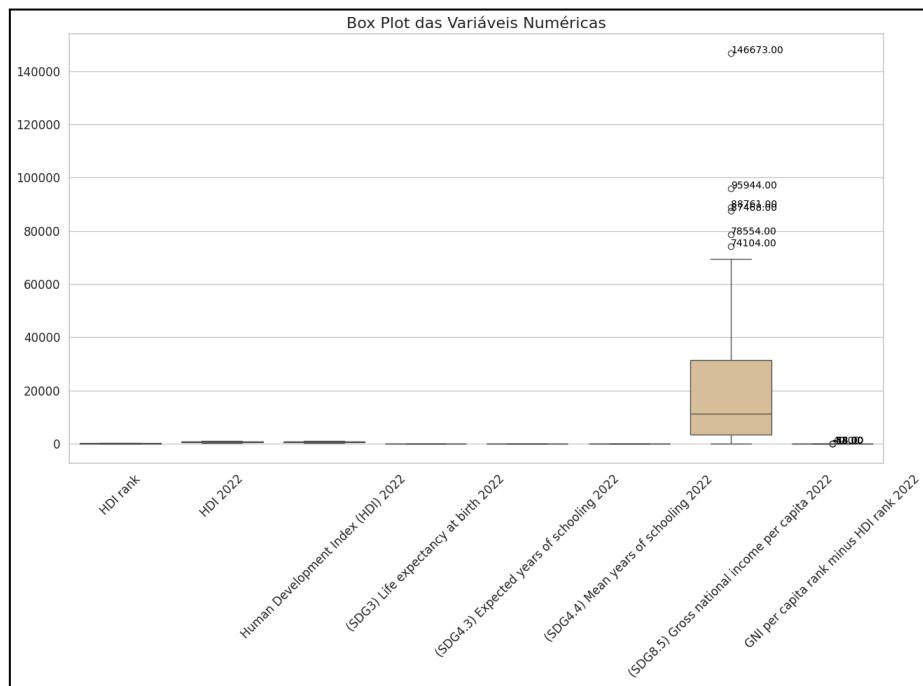
B. Box Plot



Relembreamos que o box plot, também conhecido como diagrama de caixa, é uma ferramenta gráfica que permite a visualização da distribuição de um conjunto de dados de maneira resumida e informativa, exibindo a mediana, os quartis (Q1 e Q3), os valores mínimos e máximos (sem outliers), e os outliers.

Porém, ao aplicar box plots ao dataset em questão, observou-se uma grande disparidade entre as variáveis em termos de escala, especificamente, enquanto a maioria das variáveis possui valores muito baixos e próximos uns dos outros, a variável "Gross national income per capita 2022" apresenta valores significativamente mais altos, dominando a visualização. Isso torna a comparação entre variáveis menos eficaz, pois a escala de "Gross

"national income per capita 2022" distorce a percepção das outras variáveis. Dessa forma, a análise dos box plots revelou que a maioria das variáveis tem interquartis pequenos e próximos ao eixo x, indicando baixa variação entre elas. Esse fenômeno pode ser menos informativo para identificar diferenças significativas entre essas variáveis, especialmente no contexto deste dataset. A baixa variação sugere que as distribuições das variáveis são bastante homogêneas, limitando a capacidade dos box plots de destacar discrepâncias importantes.



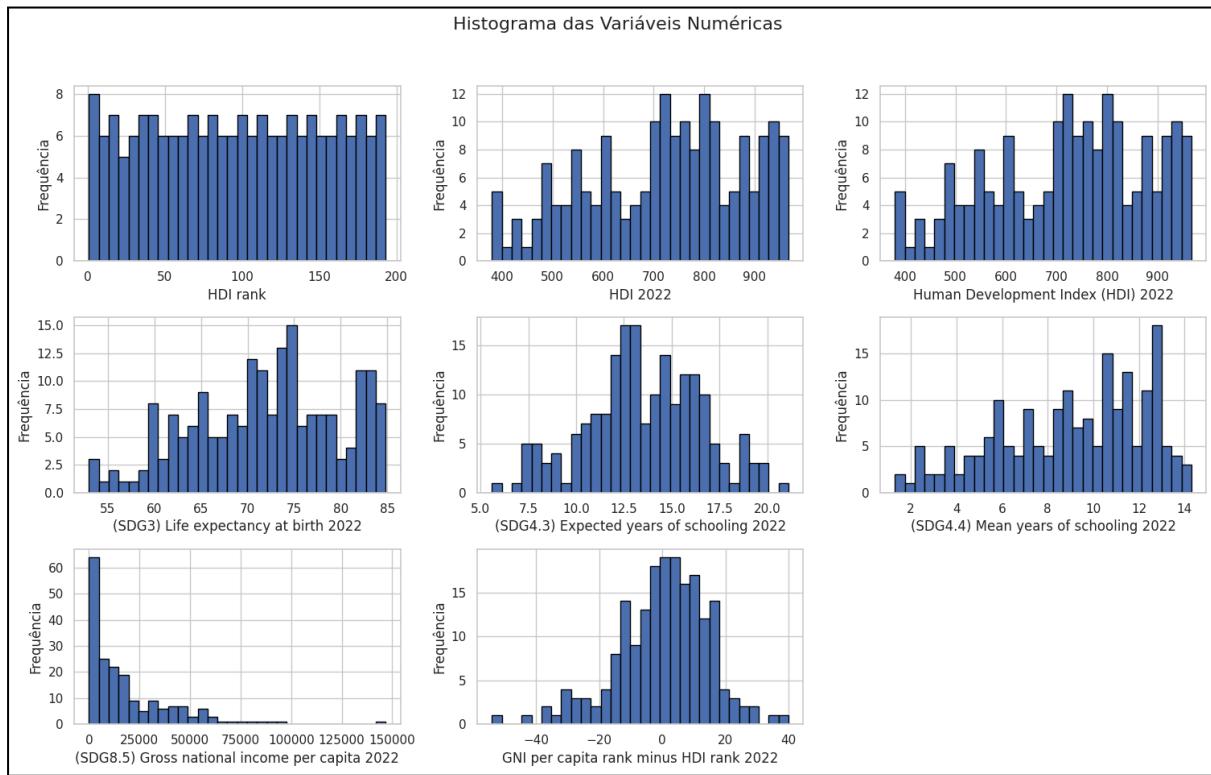
Assim, tentamos adaptar o box plot, a fim de tentar melhorar a visualização dos dados, ajustando as escalas, removendo outliers ou focando em subgrupos específicos de dados, para ver se tais adaptações poderiam fornecer uma visão mais clara e informativa da distribuição das variáveis no dataset. Porém, evidenciou-se que o box plot não seria a melhor ferramenta para adentrar e aprofundar nas análises, razão pela qual não foi utilizada mais adiante.

C. Histogramas específicos (por coluna)

Para complementar a análise, foi criada uma visualização mais direcionada para os histogramas, de forma que fossem analisadas cada coluna numérica do dataset individualmente, o que permitiu uma análise mais detalhada de cada variável individualmente. No contexto das variáveis do dataset (como HDI rank, HDI 2000, HDI

2010, etc.), a criação de histogramas individuais para cada coluna foi fundamental para entender como cada variável é distribuída.

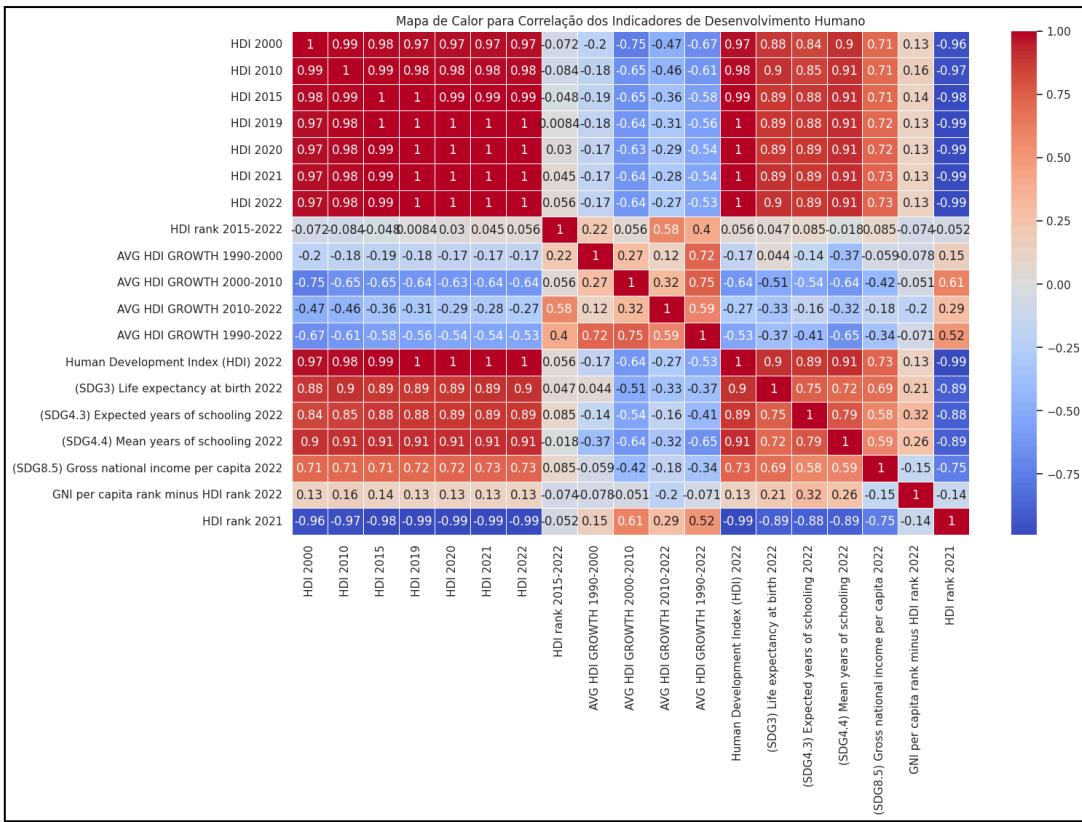
Ao analisar os histogramas das variáveis de IDH e SDG (Objetivos de Desenvolvimento Sustentável), foi possível observar tendências globais. Por exemplo, ao visualizar a distribuição do índice de desenvolvimento humano ao longo dos anos, foi possível identificar como ele evoluiu e se existem padrões consistentes.



A tentativa de adaptar a visualização de histogramas específicos por coluna foi de extrema importância para melhorar a compreensão da distribuição dos dados no dataset, tendo em vista que forneceram uma visão detalhada da frequência dos valores de cada variável.

D. Mapa de Calor

Conforme demonstrado em sala, um mapa de calor é uma representação gráfica que utiliza cores para mostrar a intensidade de valores dentro de uma matriz e, no contexto de análise de dados, os mapas de calor são frequentemente usados para visualizar a correlação entre diferentes variáveis em um dataset, permitindo uma rápida identificação de padrões e relações entre as variáveis, facilitando a interpretação dos dados.



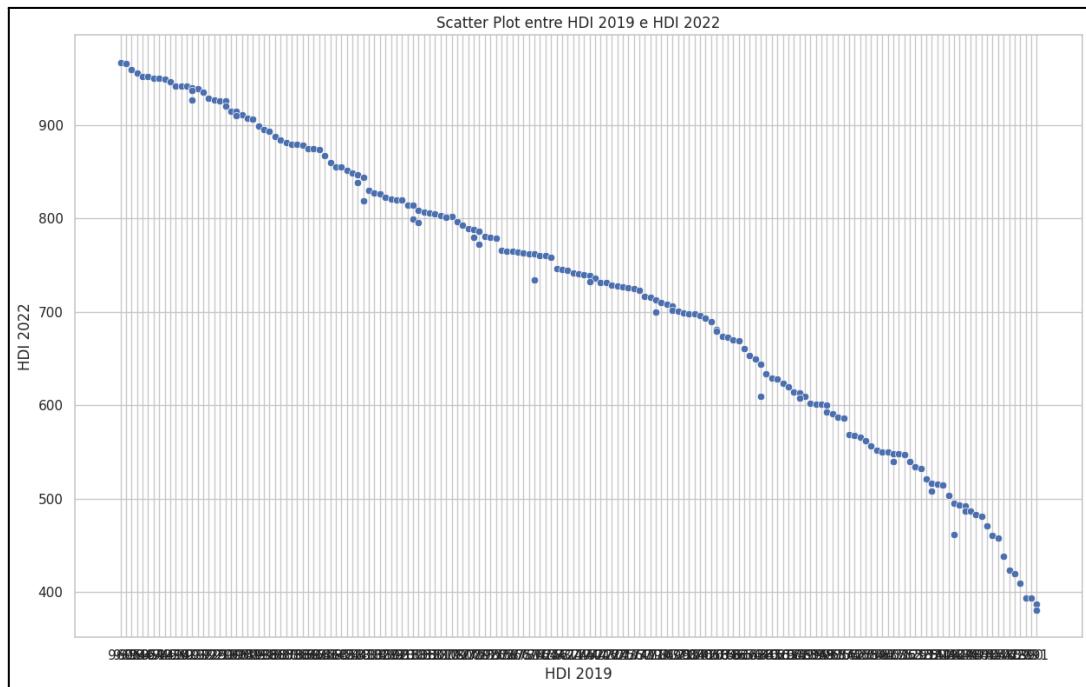
Sendo assim, verificamos que as colunas do IDH para diferentes anos mostram correlações muito altas entre si (quase 1). Isso indica que os países que tinham um IDH alto em um ano tendem a manter um IDH alto nos anos seguintes, evidenciando, ao longo do tempo, a estabilidade e o desenvolvimento contínuo dos países. Ainda, viu-se que existe uma correlação significativa entre o IDH e vários índices de SDG, demonstrando que melhorias nessas áreas estão fortemente associadas a um IDH mais alto, evidenciando a interconexão entre desenvolvimento humano e metas de sustentabilidade.

E. Scatter Plot

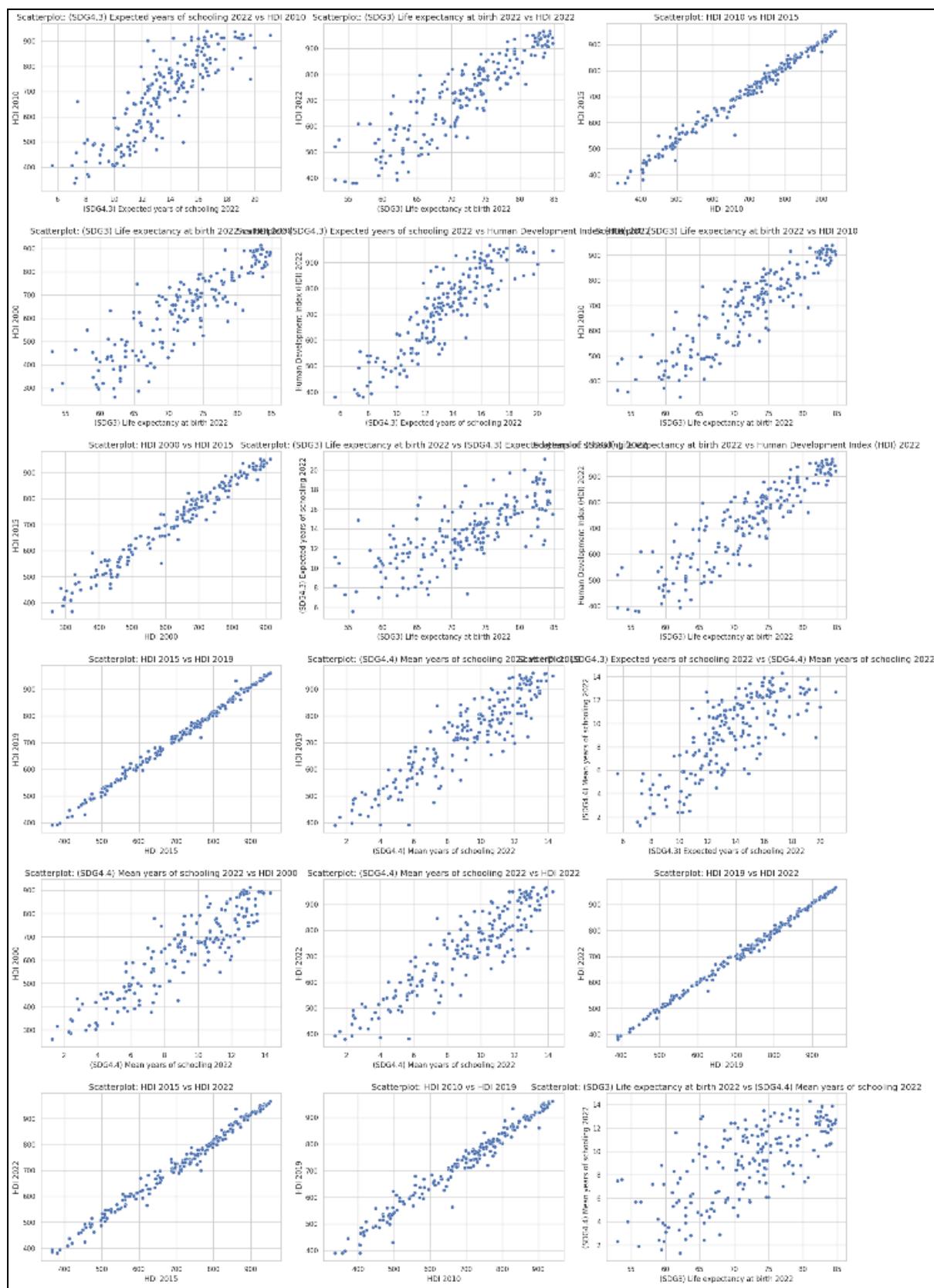
Um scatter plot, ou gráfico de dispersão, é um tipo de gráfico que utiliza pontos para representar os valores de duas variáveis diferentes., em que cada ponto representa uma observação do conjunto de dados, sendo utilizado para visualizar a relação entre duas variáveis, permitindo identificar padrões, tendências e possíveis correlações.

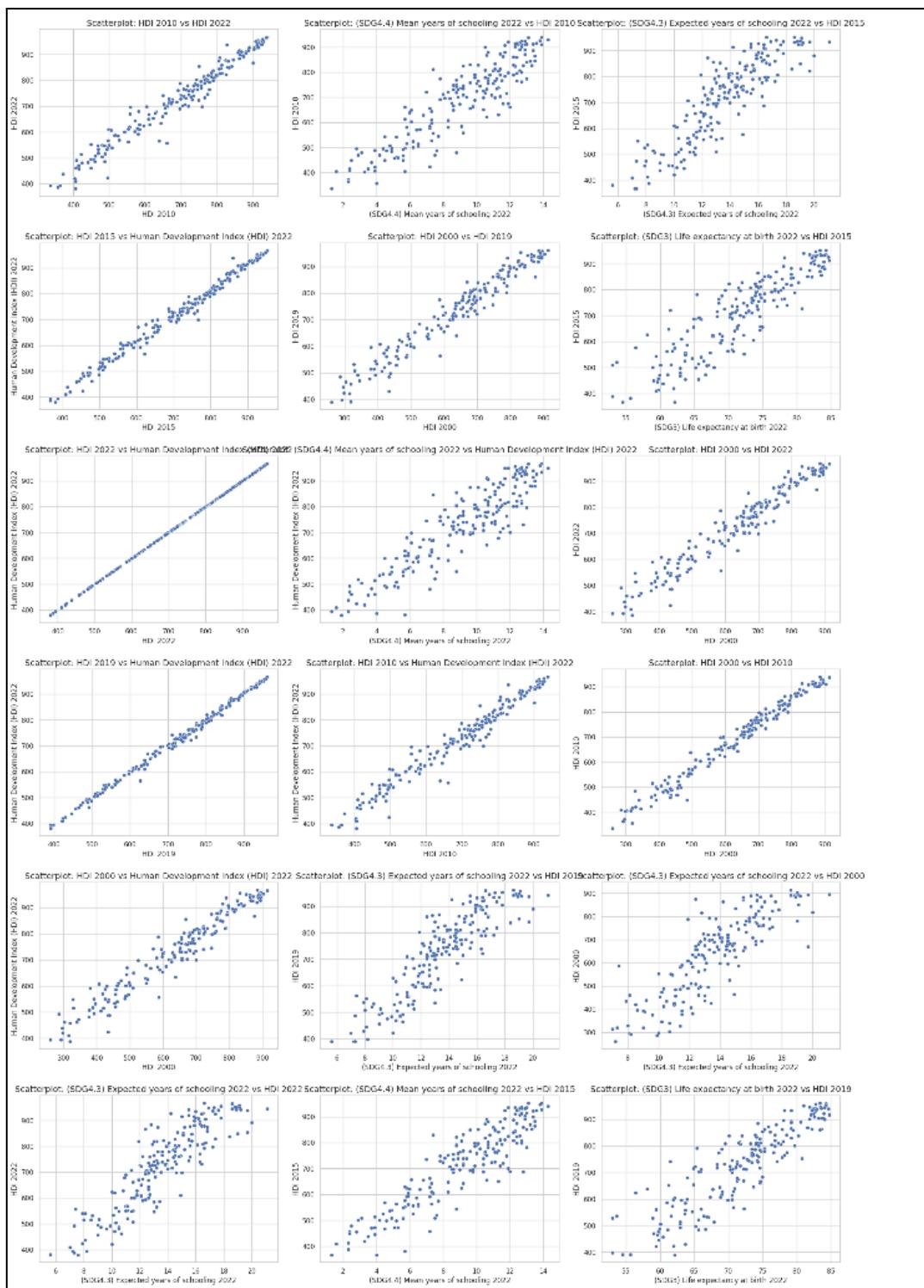
Sendo assim, o scatter plot criado entre as colunas do HDI 2019 e HDI 2022, ambos selecionados com base no visualizado via o mapa de calor acima disposto, que indicou uma alta correlação entre essas duas variáveis, sugerindo a necessidade de um aprofundamento.

Dessa forma, confirmou a alta correlação observada no mapa de calor, pois a distribuição dos pontos ao longo de uma linha reta indica uma relação linear positiva forte entre HDI 2019 e HDI 2022.



A forte relação linear positiva confirma a consistência nos níveis de desenvolvimento humano ao longo dos anos, sugerindo que os países que tinham um alto IDH em 2019 continuaram a ter altos níveis de IDH em 2022, reforçando os achados do mapa de calor e fornece uma base visual sólida para a análise de correlação entre os anos de IDH. Ainda, realizamos a correlação de todas as colunas entre si, para verificar se alguma outra tendência seria evidenciada.





Portanto, após a análise dos gráficos, verificamos que a hipótese levantada no mapa de calor, sobre a correlação entre presença de indicadores de SDG, foi confirmada, tendo em vista as altas variações vistas no IDH quando o SDG oscilava, demonstrando uma clara relação de causalidade e dependência.

3.2. Normalização

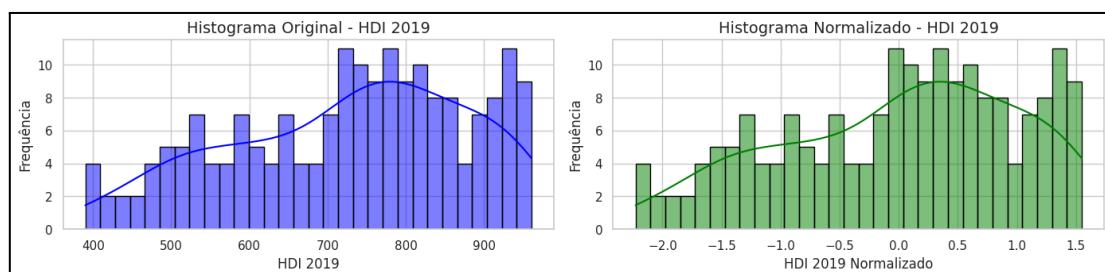
A. Normalização (Parte Matemática)

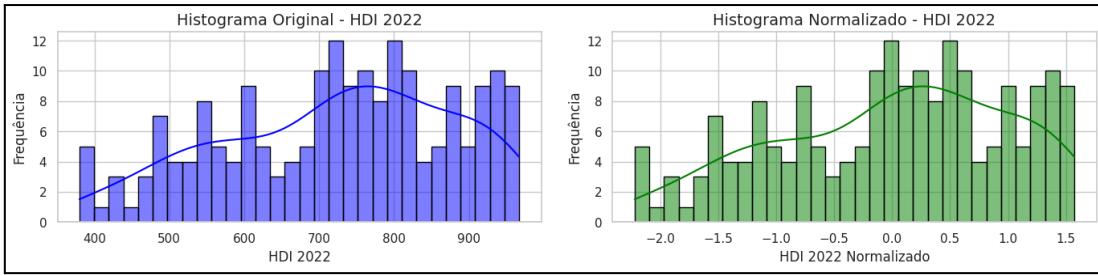
Conforme visto no semestre, a normalização dos dados é um passo crucial na análise de dados e modelagem preditiva, pois visa ajustar os valores de diferentes variáveis para uma escala comum, transformando os dados para que eles tenham uma média de 0 e um desvio padrão de 1. A fim de continuar a análise realizada na Estatística Descritiva, escolhemos as colunas 'HDI 2019' e 'HDI 2022' para normalização.

Dessa forma, a normalização dos dados das colunas 'HDI 2019' e 'HDI 2022' foi uma etapa essencial para preparar os dados para análises mais precisas e para garantir que os modelos de aprendizado de máquina pudessem ser aplicados de maneira eficaz, assegurando que todas as variáveis estejam na mesma escala, melhorando a interpretação e a comparabilidade dos dados, além de potencialmente aumentar a performance de modelos preditivos.

	HDI 2019	HDI 2022
0	1.542198	1.571697
1	1.548813	1.565236
2	1.528968	1.520009
3	1.495894	1.500626
4	1.449590	1.474782

Ainda, para melhorar a visualização da normalização realizada, foram gerados gráficos, que demonstram que apenas as escalas foram ajustadas, não prejudicando os indicativos dos dados ali dispostos.





B. Dados Matemáticos

Ainda, para melhor preparar e embasar as análises realizadas, foram gerados estatísticas descritivas da coluna 'AVG HDI GROWTH 1990-2022', agrupadas por país, incluindo medidas como contagem, média, desvio padrão, valores mínimos, quartis e valores máximos. Ressaltamos que as estatísticas descritivas fornecem uma visão abrangente das características dos dados, incluindo tendência central, dispersão e distribuição.

Ao agrupar por país e calcular essas estatísticas ajuda a identificar padrões e diferenças no crescimento médio do IDH ao longo do período de 1990 a 2022 entre diferentes países, é possível visualizar e mapear políticas públicas e decisões estratégicas, evidenciando quais países estão progredindo mais rapidamente e quais podem precisar de mais suporte. Abaixo, apresentamos um recorte do resultado das operações.

	count	mean	std	min	25%	50%	75%	max
Country								
Afghanistan	1.0	1.53	NaN	1.53	1.53	1.53	1.53	1.53
Albania	1.0	0.61	NaN	0.61	0.61	0.61	0.61	0.61
Algeria	1.0	0.72	NaN	0.72	0.72	0.72	0.72	0.72
Argentina	1.0	0.50	NaN	0.50	0.50	0.50	0.50	0.50
Armenia	1.0	0.56	NaN	0.56	0.56	0.56	0.56	0.56
...
Venezuela (Bolivarian Republic of)	1.0	0.19	NaN	0.19	0.19	0.19	0.19	0.19
Viet Nam	1.0	1.22	NaN	1.22	1.22	1.22	1.22	1.22
Yemen	1.0	0.54	NaN	0.54	0.54	0.54	0.54	0.54
Zambia	1.0	0.98	NaN	0.98	0.98	0.98	0.98	0.98
Zimbabwe	1.0	0.43	NaN	0.43	0.43	0.43	0.43	0.43

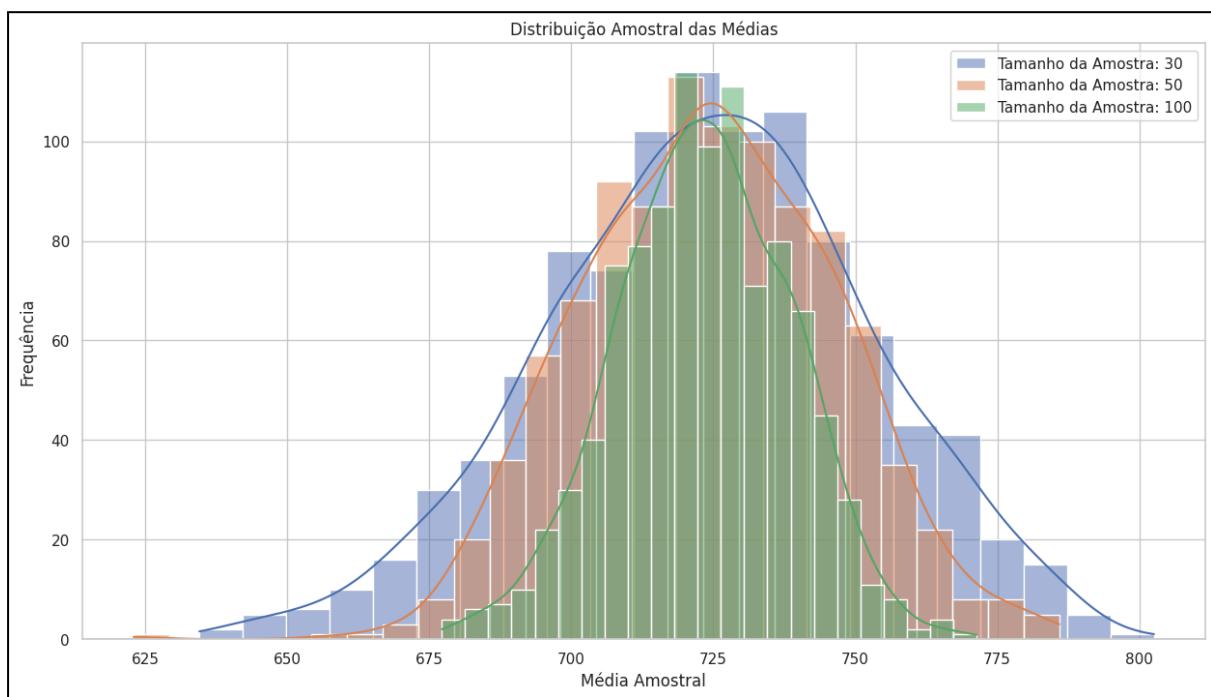
143 rows x 8 columns

3.3. Distribuição Amostral

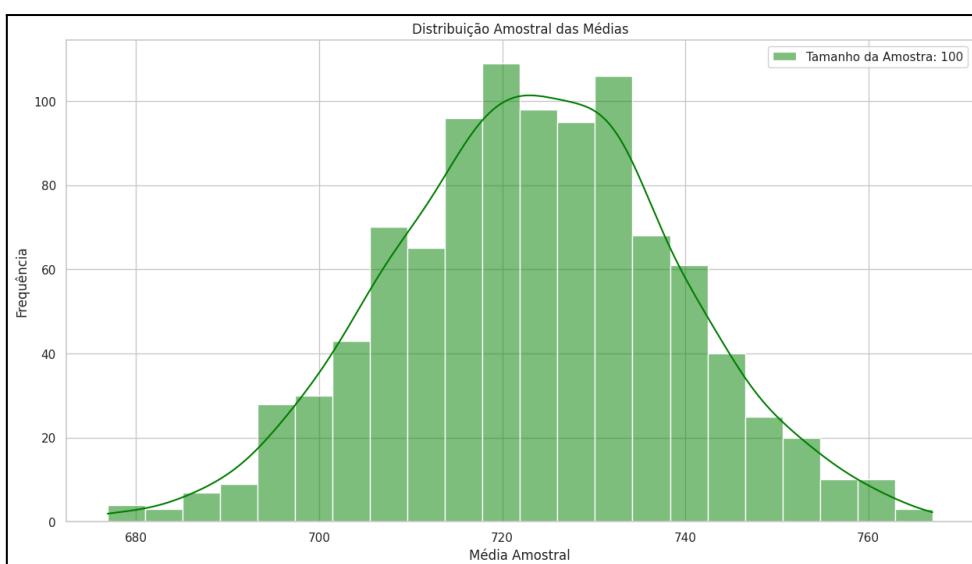
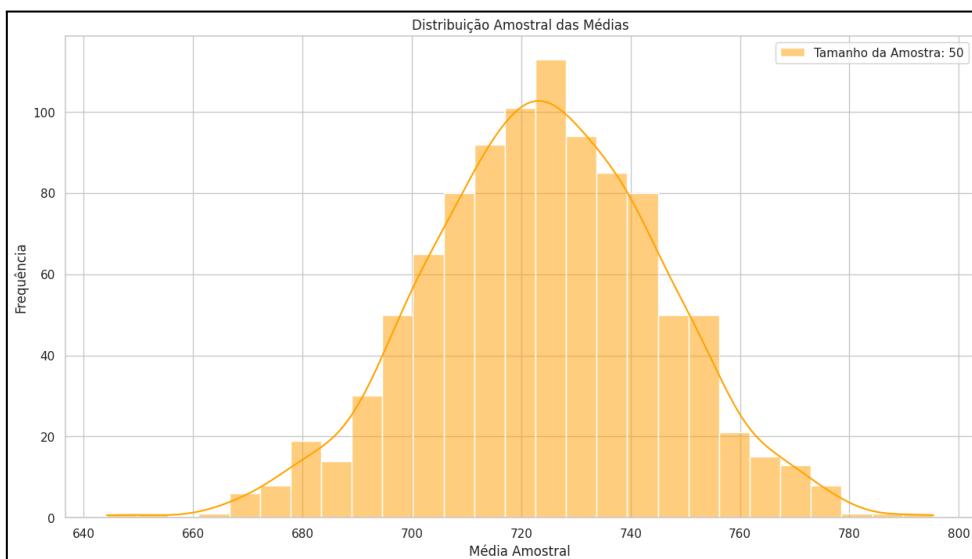
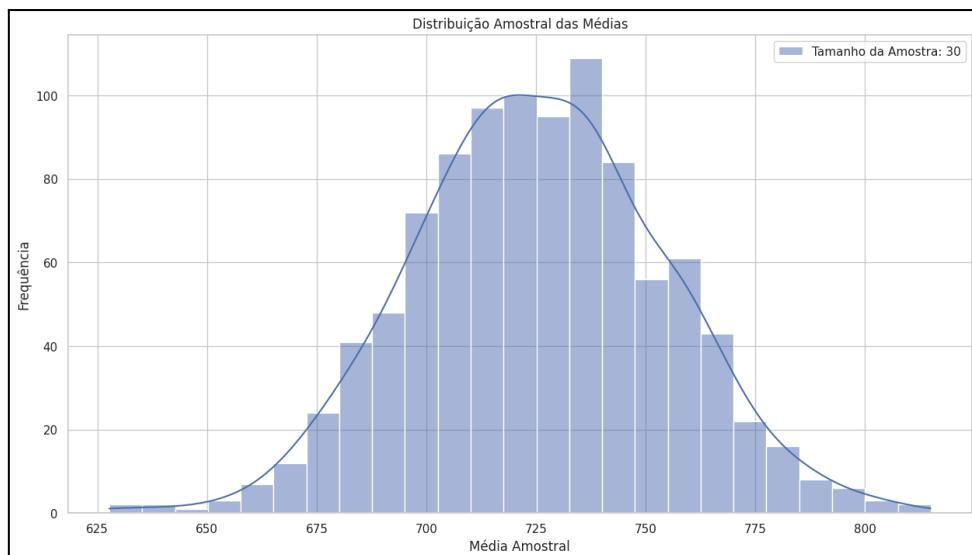
Prosseguindo, vejamos que a distribuição amostral refere-se à distribuição de uma estatística (como a média, mediana, variância) calculada a partir de múltiplas amostras de um

conjunto de dados. Em outras palavras, ela descreve como uma estatística varia de amostra para amostra dentro de uma população, sendo fundamental para a inferência estatística, pois permite fazer estimativas sobre a população com base em amostras. Dessa forma, a distribuição amostral permite fazer inferências sobre a população inteira a partir de uma ou mais amostras, fornecendo uma base para estimar parâmetros populacionais, como a média e a variância, a partir das estatísticas da amostra.

Em nossa análise de dados, a distribuição amostral foi visualizada e analisada para calcular a média do crescimento do IDH em várias amostras de países, no qual utilizamos para construir intervalos de confiança para a média do crescimento do IDH.



Conforme demonstrado acima, utilizamos como tamanhos de amostra os valores de 30, 50 e 100, cujo gráficos individuais estão abaixo disponibilizados.



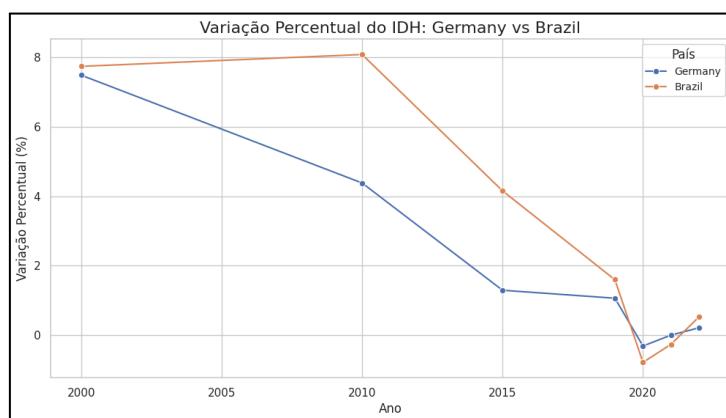
Com base no gráfico de distribuição amostral das médias, vimos que os diferentes tamanhos de amostra (30, 50 e 100) parecem se aproximar de uma distribuição normal, como evidenciado pelas curvas de densidade suavizadas sobre os histogramas. Além disso, à medida que o tamanho da amostra aumenta (de 30 para 50 e 100), a média amostral tende a se concentrar mais próxima da média populacional, conforme indicado pela concentração das distribuições ao redor de um ponto central (aproximadamente entre 700 e 725).

Vejamos que a forma das distribuições se aproximando de uma curva normal, independentemente do tamanho da amostra, é uma manifestação do Teorema do Limite Central, que diz que a soma (ou média) de um grande número de variáveis aleatórias independentes e identicamente distribuídas tende a ser distribuída normalmente, independentemente da distribuição original das variáveis.

A. Line Plot (Comparação entre países)

Para aprofundar ainda mais a análise, decidimos trazer a pauta de comparar diferentes países no contexto de indicadores de desenvolvimento humano. No exemplo proposto, temos a Alemanha (um país desenvolvido) e Brasil (um país em desenvolvimento), a fim de evidenciar se, dado suas trajetórias de desenvolvimento diferentes, comparar os dois pode ajudar a entender como diferentes contextos socioeconômicos, políticas e intervenções impactam o desenvolvimento humano.

Enquanto países desenvolvidos como a Alemanha podem ter implementado políticas e programas de sucesso que podem ser adaptados e aplicados em contextos semelhantes em países em desenvolvimento como o Brasil. A análise das variações percentuais no IDH ao longo do tempo pode ajudar a identificar períodos de melhorias significativas e correlacionar esses períodos com políticas específicas.



3.4. Intervalo de Confiança

Chegando ao fim das análises propostas, vejamos que o intervalo de confiança é uma ferramenta estatística fundamental que proporciona insights importantes sobre a precisão e a confiabilidade das estimativas de parâmetros populacionais. Tendo em vista que fornece um intervalo dentro do qual se espera que o verdadeiro valor do parâmetro populacional (como a média) se encontra, foi essencial para ajudar a entender a precisão da estimativa obtida a partir da amostra. Dessa forma, ele reflete a variabilidade dos dados e o grau de incerteza associado à estimativa, pois um intervalo de confiança mais estreito indica menor variabilidade e maior precisão, enquanto um intervalo mais amplo sugere maior variabilidade e menor precisão.

Partindo disso, calculamos a verdadeira média populacional se encontre com os níveis de confiança especificados, abaixo disposto.

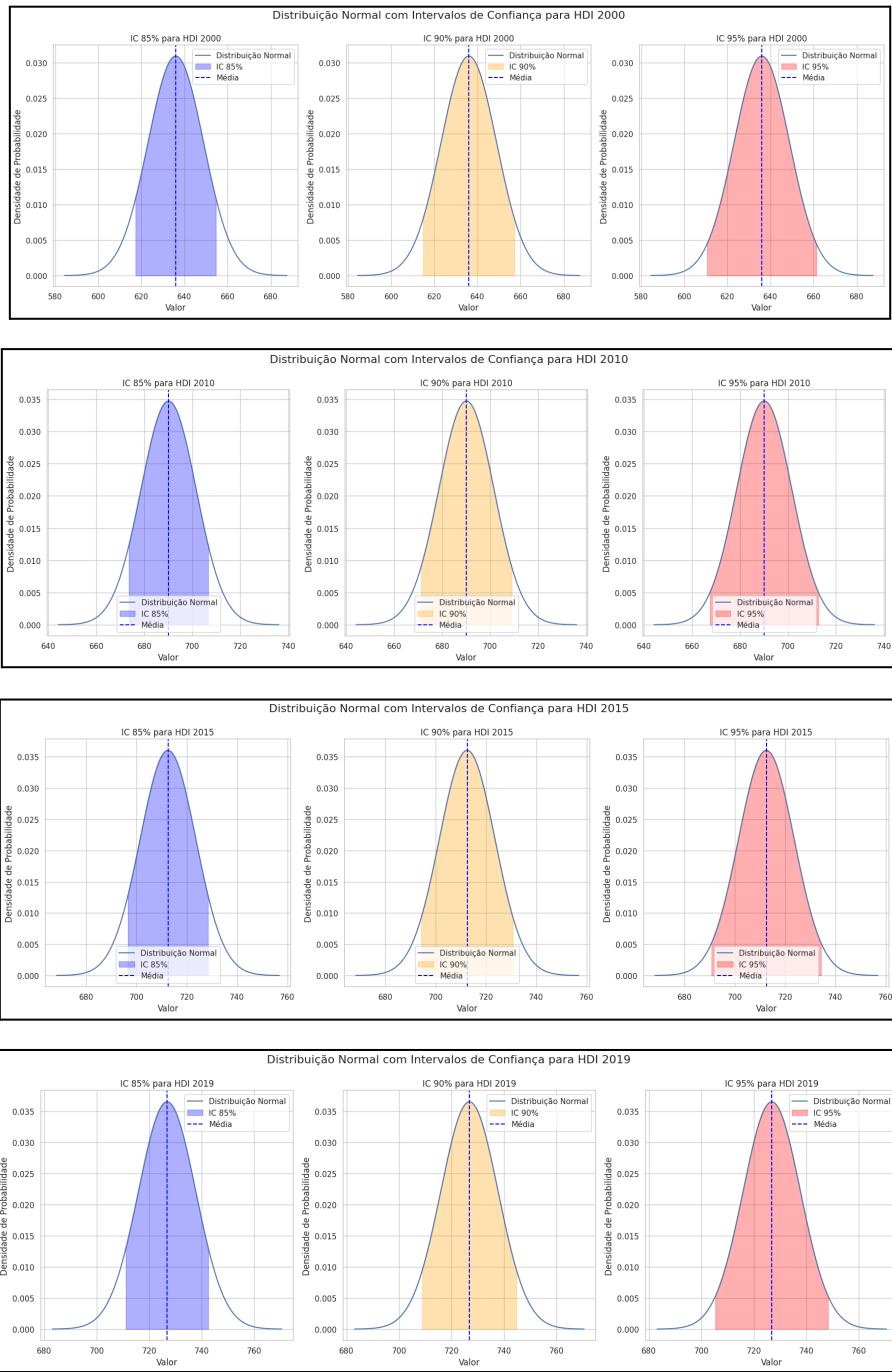
Intervalo de Confiança de 85%: (707.5969149767831, 739.8849503081909)
Intervalo de Confiança de 90%: (705.2789916599077, 742.2028736250663)
Intervalo de Confiança de 95%: (701.7094480388959, 745.7724172460781)

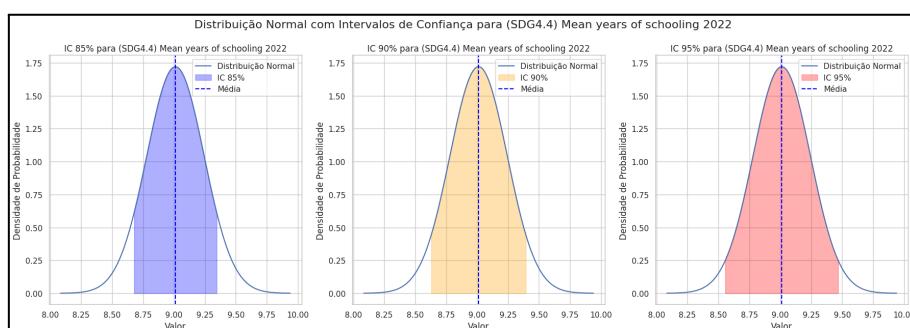
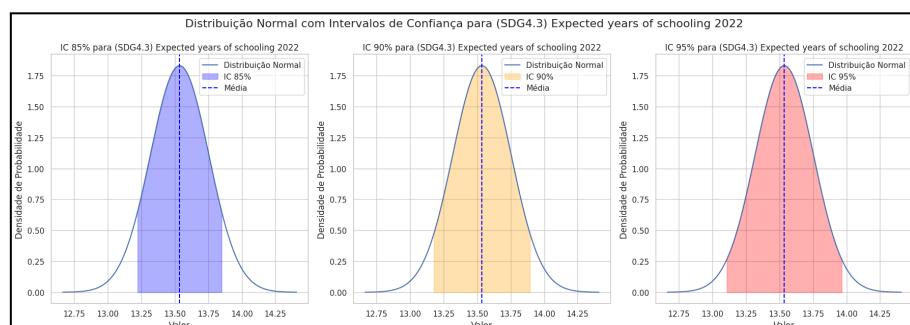
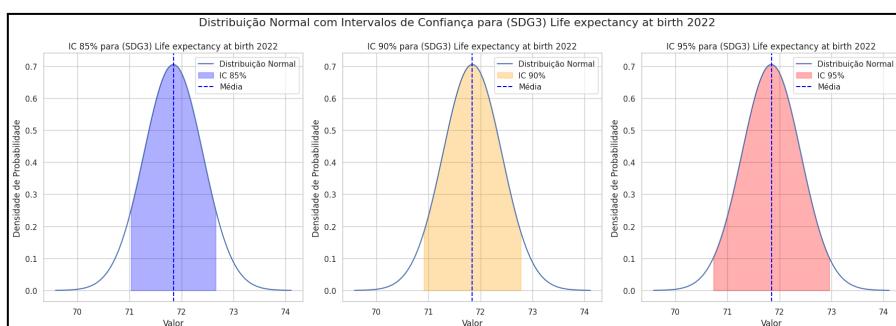
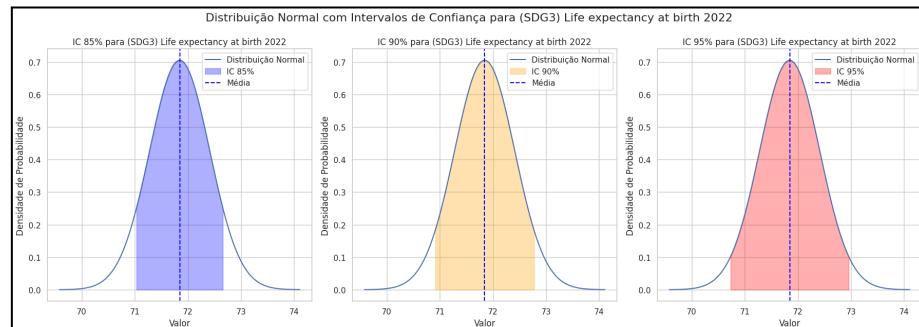
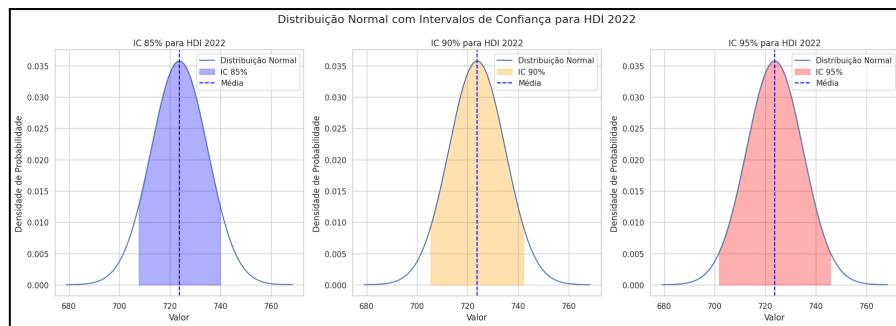
Ainda, realizamos o cálculo do intervalo de confiança para cada uma das colunas numéricas, a fim de ter isso como embasamento para análises e aprofundamentos futuros.

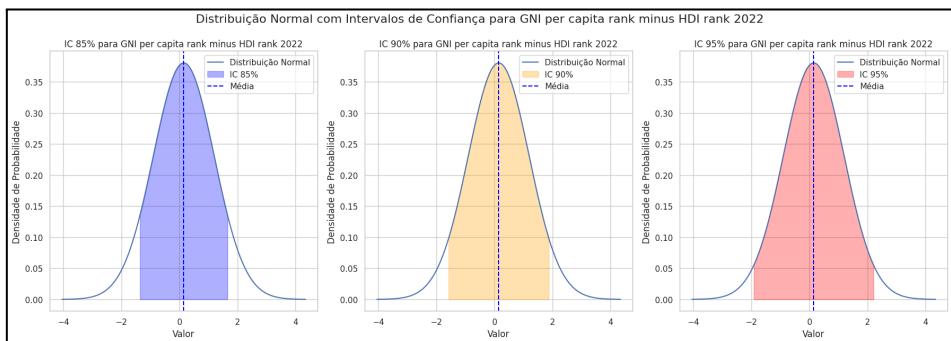
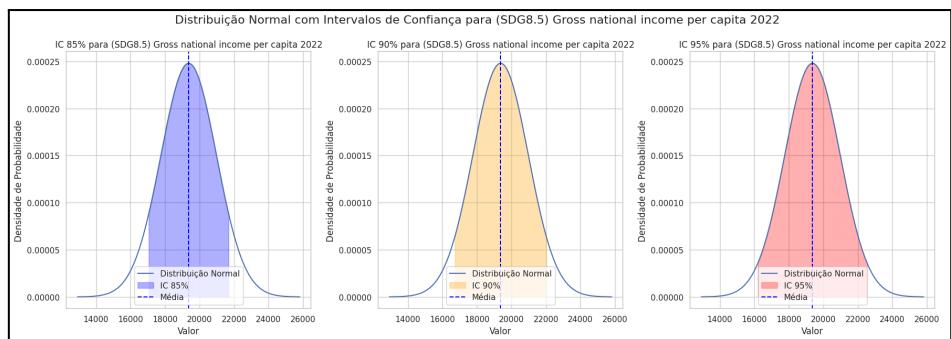
Variável: HDI rank
Média: 96.8549
Erro padrão da média: 4.0246
Intervalo de Confiança de 85%: (91.03809924675784, 102.67174531282765)
Intervalo de Confiança de 90%: (90.20293231262765, 103.50691224695784)
Intervalo de Confiança de 95%: (88.91679623534664, 104.79304832423885)
Variável: HDI 2000
Média: 635.9489
Erro padrão da média: 12.8954
Intervalo de Confiança de 85%: (617.3036599478208, 654.5940673249064)
Intervalo de Confiança de 90%: (614.6248984204535, 657.27288522737)
Intervalo de Confiança de 95%: (610.4902906808315, 661.3994285918958)
Variável: HDI 2010
Média: 690.0838
Erro padrão da média: 11.4880
Intervalo de Confiança de 85%: (673.4792965590001, 706.6882427080157)
Intervalo de Confiança de 90%: (671.0950958149006, 709.0724434521151)
Intervalo de Confiança de 95%: (667.42335411839, 712.7441851551766)
Variável: HDI 2015
Média: 712.52609
Erro padrão da média: 11.0751
Intervalo de Confiança de 85%: (696.5187027364148, 728.5333805969185)
Intervalo de Confiança de 90%: (694.2263237535988, 730.8317595798235)
Intervalo de Confiança de 95%: (690.6800143803743, 734.371268952959)
Variável: HDI 2019
Média: 726.8594
Erro padrão da média: 10.9386
Intervalo de Confiança de 85%: (711.0493579564533, 742.6693920435467)
Intervalo de Confiança de 90%: (708.7793110079513, 744.939438920487)
Intervalo de Confiança de 95%: (705.2834330385463, 748.4353169614537)
Variável: HDI 2022
Média: 723.7409
Erro padrão da média: 11.1699
Intervalo de Confiança de 85%: (707.5969149767831, 739.8849503081909)
Intervalo de Confiança de 90%: (705.2789916599077, 742.2028736250663)
Intervalo de Confiança de 95%: (701.7094480388959, 745.7724172460781)
Variável: AVG HDI GROWTH 1990-2022
Média: 0.7188
Erro padrão da média: 0.0358
Intervalo de Confiança de 85%: (0.6670138781962573, 0.77060849942612)
Intervalo de Confiança de 90%: (0.6595595583218367, 0.7780628193065406)
Intervalo de Confiança de 95%: (0.6480662193989121, 0.7895561582234653)

Variável: (SDG3) Life expectancy at birth 2022
Média: 71.8446
Erro padrão da média: 0.5659
Intervalo de Confiança de 85%: (71.02667871528372, 72.66244045570075)
Intervalo de Confiança de 90%: (70.90924913989927, 72.7798700310852)
Intervalo de Confiança de 95%: (70.72841054671548, 72.96070862426899)
Variável: (SDG4.3) Expected years of schooling 2022
Média: 13.5358
Erro padrão da média: 0.2180
Intervalo de Confiança de 85%: (13.220616227252435, 13.850886363421136)
Intervalo de Confiança de 90%: (13.17536981230149, 13.896132778372081)
Intervalo de Confiança de 95%: (13.105691471052689, 13.965811119620883)
Variável: (SDG4.4) Mean years of schooling 2022
Média: 9.0119
Erro padrão da média: 0.2319
Intervalo de Confiança de 85%: (8.67677583598234, 9.347058360908854)
Intervalo de Confiança de 90%: (8.628656974692877, 9.39517722198317)
Intervalo de Confiança de 95%: (8.554555138313887, 9.469279058577307)
Variável: (SDG8.5) Gross national income per capita 2022
Média: 19370.5434
Erro padrão da média: 1609.9733
Intervalo de Confiança de 85%: (17043.62710602166, 21697.459629729638)
Intervalo de Confiança de 90%: (16709.533476328354, 22031.553259422944)
Intervalo de Confiança de 95%: (16195.037687787773, 22546.049047963523)
Variável: GNI per capita rank minus HDI rank 2022
Média: 0.1451
Erro padrão da média: 1.0492
Intervalo de Confiança de 85%: (-1.371343949063297, 1.6614993894778054)
Intervalo de Confiança de 90%: (-1.5890685092193741, 1.8792239496338814)
Intervalo de Confiança de 95%: (-1.9243588569131713, 2.2145142973276792)

Abaixo, disponibilizamos graficamente os intervalos de confiança relevantes, evidenciados pelos cálculos acima realizados.







4. ANÁLISES APROFUNDADAS

Conforme dito anteriormente, baseando-me nas colunas do dataset, que incluem rankings de Índice de Desenvolvimento Humano (IDH) - descrito no dataset no inglês, na forma do Human Development Index (HDI) -, seus valores em diferentes anos, expectativa de vida, anos esperados de escolaridade, anos médios de escolaridade, e renda nacional bruta per capita, entre outros, visualizamos as algumas das possíveis relações e inferências entre e sobre esses dados, conforme processo descrito abaixo.

Antes de prosseguir, foi criado um mapeamento para associar cada país do dataset ao seu respectivo continente, adicionando uma nova coluna ao dataset chamada 'Continent', em que cada país foi associado ao seu continente correspondente com base no mapeamento estabelecido.

```
continent_mapping = {
    'Afghanistan': 'Asia',
    'Albania': 'Europe',
    'Algeria': 'Africa',
    'Andorra': 'Europe',
    'Angola': 'Africa',
    'Antigua and Barbuda': 'North America',
    'Argentina': 'South America',
    'Armenia': 'Asia',
    'Australia': 'Oceania',
    'Austria': 'Europe',
    'Azerbaijan': 'Asia',
    'Bahamas': 'North America',
    'Bahrain': 'Asia',
    'Bangladesh': 'Asia',
    'Barbados': 'North America',
    'Belarus': 'Europe',
    'Belgium': 'Europe',
    'Belize': 'North America',
    'Benin': 'Africa',
    'Bhutan': 'Asia',
    'Bolivia': 'South America',
    'Bosnia and Herzegovina': 'Europe',
    'Botswana': 'Africa',
    'Brazil': 'South America',
    'Brunei Darussalam': 'Asia',
    'Bulgaria': 'Europe',
    'Burkina Faso': 'Africa',
    'Burundi': 'Africa',
    'Cabo Verde': 'Africa',
    'Cambodia': 'Asia',
    'Cameroon': 'Africa',
    'Canada': 'North America',
    'Central African Republic': 'Africa',
    'Chad': 'Africa',
    'Chile': 'South America',
    'China': 'Asia',
    'Colombia': 'South America',
    'Comoros': 'Africa',
    'Congo (Democratic Republic of the)': 'Africa',
    'Congo': 'Africa',
    'Costa Rica': 'North America',
    'Croatia': 'Europe',
    'Cuba': 'North America',
    'Cyprus': 'Asia',
    'Czech Republic': 'Europe',
    'Denmark': 'Europe',
    'Djibouti': 'Africa',
    'Dominica': 'North America',
    'Dominican Republic': 'North America',
    'Ecuador': 'South America',
    'Egypt': 'Africa',
    'El Salvador': 'North America',
    'Equatorial Guinea': 'Africa',
    'Eritrea': 'Africa',
    'Estonia': 'Europe',
    'Eswatini': 'Africa',
    'Ethiopia': 'Africa',
    'Fiji': 'Oceania',
    'Finland': 'Europe',
    'France': 'Europe',
    'Gabon': 'Africa',
    'Gambia': 'Africa',
    'Georgia': 'Asia',
    'Germany': 'Europe',
    'Ghana': 'Africa',
    'Greece': 'Europe',
    'Grenada': 'North America',
    'Guatemala': 'North America',}
```

```
'Guinea': 'Africa',
'Guinea-Bissau': 'Africa',
'Guyana': 'South America',
'Haiti': 'North America',
'Honduras': 'North America',
'Hong Kong, China (SAR)':
'Asia',
'Hungary': 'Europe',
'Iceland': 'Europe',
'India': 'Asia',
'Indonesia': 'Asia',
'Iran (Islamic Republic of)': 'Asia',
'Iraq': 'Asia',
'Ireland': 'Europe',
'Israel': 'Asia',
'Italy': 'Europe',
'Jamaica': 'North America',
'Japan': 'Asia',
'Jordan': 'Asia',
'Kazakhstan': 'Asia',
'Kenya': 'Africa',
'Kiribati': 'Oceania',
'Kuwait': 'Asia',
'Kyrgyzstan': 'Asia',
'Lao People\\s Democratic
Republic': 'Asia',
'Latvia': 'Europe',
'Lebanon': 'Asia',
'Lesotho': 'Africa',
'Liberia': 'Africa',
'Libya': 'Africa',
'Liechtenstein': 'Europe',
'Lithuania': 'Europe',
'Luxembourg': 'Europe',
'Madagascar': 'Africa',
'Malawi': 'Africa',
'Malaysia': 'Asia',
'Maldives': 'Asia',
'Mali': 'Africa',
'Malta': 'Europe',
'Marshall Islands': 'Oceania',
'Mauritania': 'Africa',
'Mauritius': 'Africa',
'Mexico': 'North America',
'Micronesia (Federated States
of)': 'Oceania',
'Moldova (Republic of)':
'Europe',
'Monaco': 'Europe',
'Mongolia': 'Asia',
'Montenegro': 'Europe',
'Morocco': 'Africa',
'Mozambique': 'Africa',
'Myanmar': 'Asia',
'Namibia': 'Africa',
'Nauru': 'Oceania',
'Nepal': 'Asia',
'Netherlands': 'Europe',
'New Zealand': 'Oceania',
'Nicaragua': 'North America',
'Niger': 'Africa',
'Nigeria': 'Africa',
'North Macedonia': 'Europe',
'Norway': 'Europe',
'Oman': 'Asia',
'Pakistan': 'Asia',
'Palau': 'Oceania',
'Panama': 'North America',
'Papua New Guinea': 'Oceania',
'Paraguay': 'South America',
'Peru': 'South America',
'Philippines': 'Asia',
'Poland': 'Europe',
'Portugal': 'Europe',
'Qatar': 'Asia',
'Republic of Korea': 'Asia',
'Romania': 'Europe',
'Russian Federation': 'Europe',
'Rwanda': 'Africa',
'Saint Kitts and Nevis': 'North
America',
'Saint Lucia': 'North America',
'Saint Vincent and the
Grenadines': 'North America',
'Samoa': 'Oceania',
'San Marino': 'Europe',
'Sao Tome and Principe':
'Asia',
'Saudi Arabia': 'Asia',
'Senegal': 'Africa',
'Serbia': 'Europe',
'Seychelles': 'Africa',
'Sierra Leone': 'Africa',
'Singapore': 'Asia',
'Slovakia': 'Europe',
'Slovenia': 'Europe',
'Solomon Islands': 'Oceania',
'Somalia': 'Africa',
'South Africa': 'Africa',
'South Sudan': 'Africa',
'Spain': 'Europe',
'Sri Lanka': 'Asia',
'Sudan': 'Africa',
'Suriname': 'South America',
'Sweden': 'Europe',
'Switzerland': 'Europe',
'Syrian Arab Republic': 'Asia',
'Tajikistan': 'Asia',
'Thailand': 'Asia',
'Timor-Leste': 'Asia',
'Togo': 'Africa',
'Tonga': 'Oceania',
'Trinidad and Tobago': 'North
America',
'Tunisia': 'Africa',
'T\\rkiye': 'Asia',
'Turkmenistan': 'Asia',
'Tuvalu': 'Oceania',
'Uganda': 'Africa',
'Ukraine': 'Europe',
'United Arab Emirates': 'Asia',
```

```

'United Kingdom': 'Europe',
'United States': 'North
America',
'Uruguay': 'South America',
'Uzbekistan': 'Asia',
'Vanuatu': 'Oceania',
'Venezuela (Bolivarian Republic
of)': 'South America',
}

'Viet Nam': 'Asia',
'Yemen': 'Asia',
'Zambia': 'Africa',
'Zimbabwe': 'Africa'
}

hdi['Continent'] =
hdi['Country'].map(continent_mappin
g)

```

Com a coluna de 'Continent', podemos comparar o IDH de países dentro da mesma região, permitindo a análise de tendências de desenvolvimento humano dentro de continentes específicos, identificando se certas regiões estão progredindo mais rapidamente do que outras e investigar os motivos por trás dessas tendências. Além disso, a coluna 'Continent' facilita a avaliação do impacto de políticas governamentais regionais e a identificação de países que são outliers dentro de seus continentes, permitindo uma análise mais aprofundada sobre os fatores que os tornam diferentes.

Assim, para a primeira análise aprofundada, olhamos a evolução do IDH ao longo do período de tempo descrito no dataset (2000 a 2022), comparando os valores do IDH em diferentes anos para observar a evolução e identificar tendências de crescimento ou declínio. Além disso, também buscamos correlacionar esses valores com os índices de Sustainable Development Goals (SDG) - no português Objetivos de Desenvolvimento Sustentável (ODS) - buscando entender como a presença ou não desses índices afetam o IDH.

Inicialmente, buscamos verificar as variações intercontinentais, visualizando como cada país se comportou no intervalo analisado, em razão da correlação com os índices de SDG estipulados. Abaixo, disponibilizamos o código utilizado:

```

years = ['HDI 2000', 'HDI 2010', 'HDI 2015', 'HDI 2019', 'HDI 2020', 'HDI
2021', 'HDI 2022']
sdg_columns = [
    '(SDG3) Life expectancy at birth 2022',
    '(SDG4.3) Expected years of schooling 2022',
    '(SDG4.4) Mean years of schooling 2022',
    '(SDG8.5) Gross national income per capita 2022'
]
selected_columns = years + sdg_columns

hdi_filtered = hdi[selected_columns]
correlation_matrix = hdi_filtered.corr()

correlated_columns = correlation_matrix[years].apply(lambda x: x >
0.8).any(axis=1)
correlated_columns = correlated_columns[correlated_columns].index.tolist()

correlated_data = hdi[['Country', 'Continent']] + correlated_columns

```

```

hdi_long = pd.melt(correlated_data, id_vars=['Country', 'Continent'],
value_vars=years, var_name='Year', value_name='HDI')

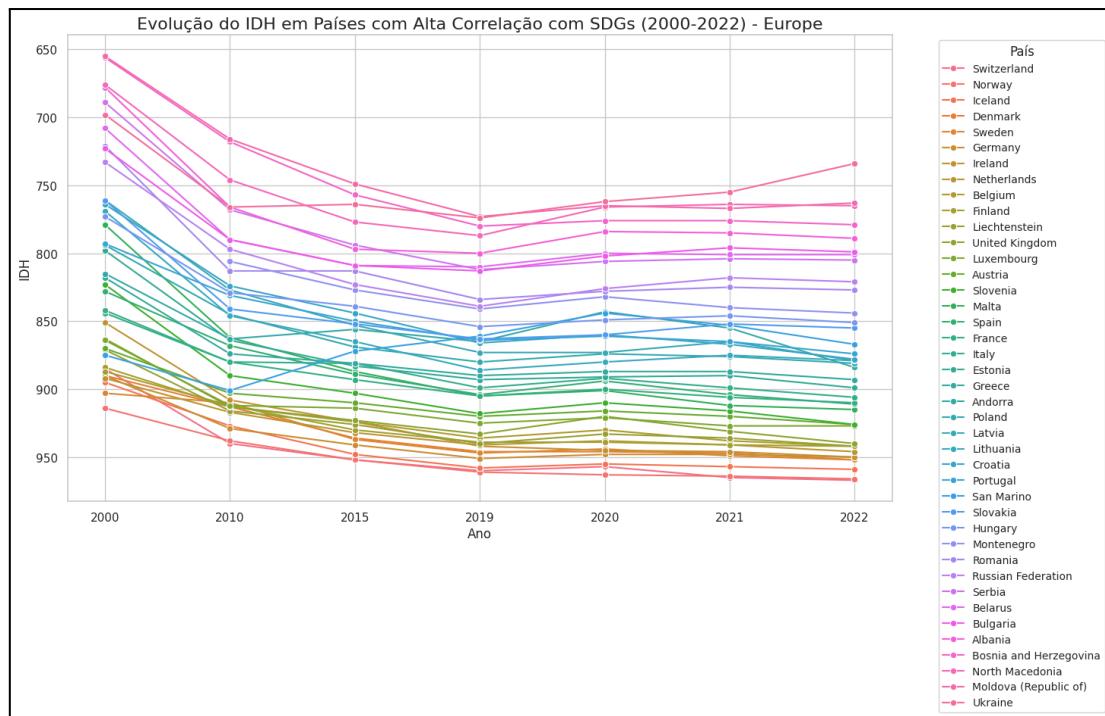
# tivemos que converter a coluna 'Year' para string para facilitar a plotagem
hdi_long['Year'] = hdi_long['Year'].str.extract(r'(\d{4})')[0]

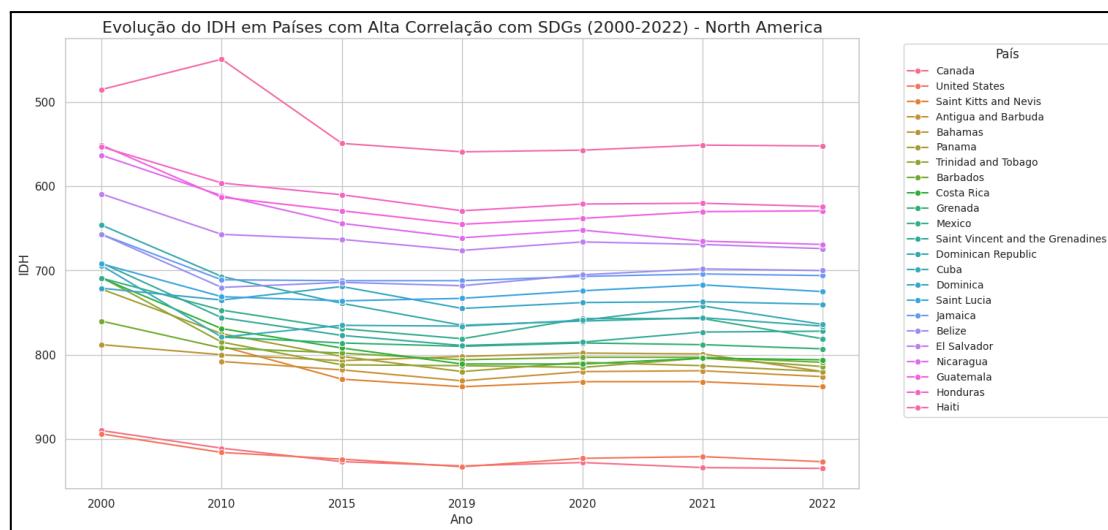
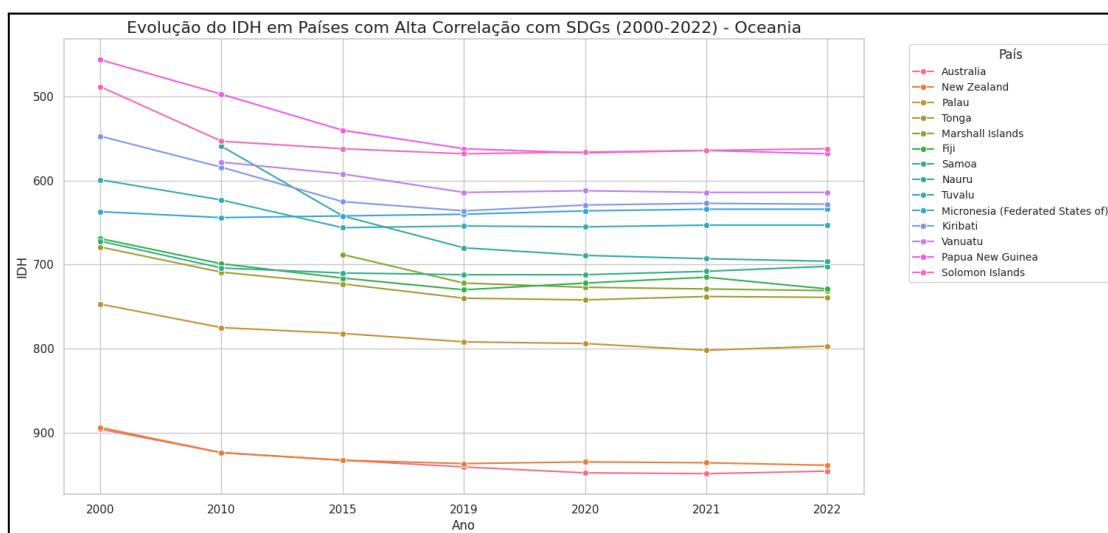
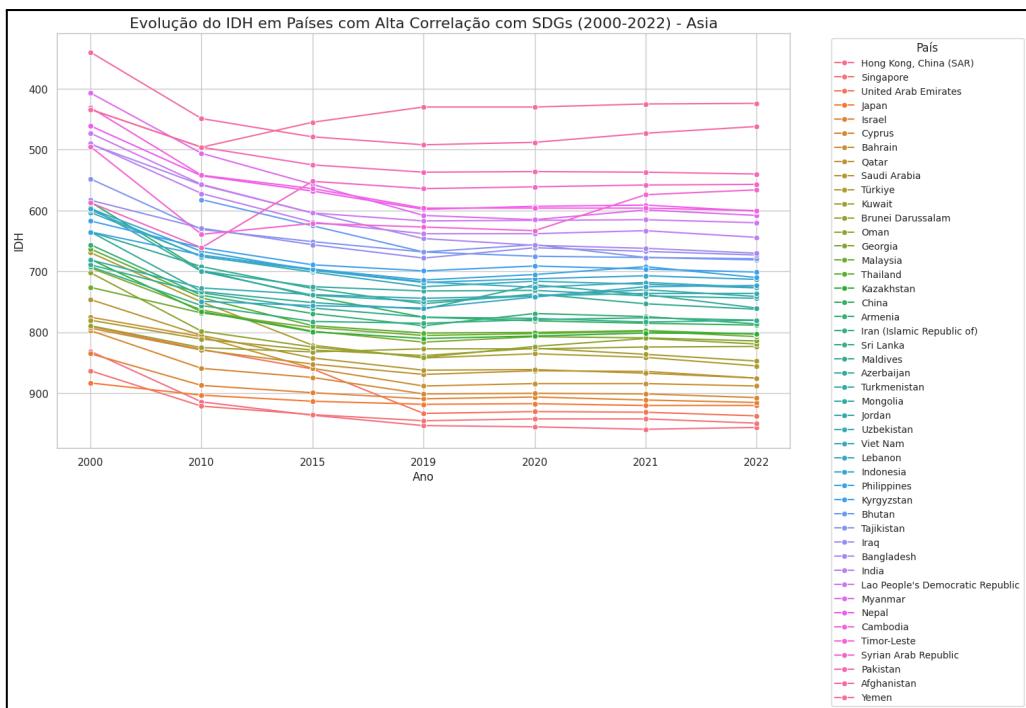
continents = hdi_long['Continent'].unique()

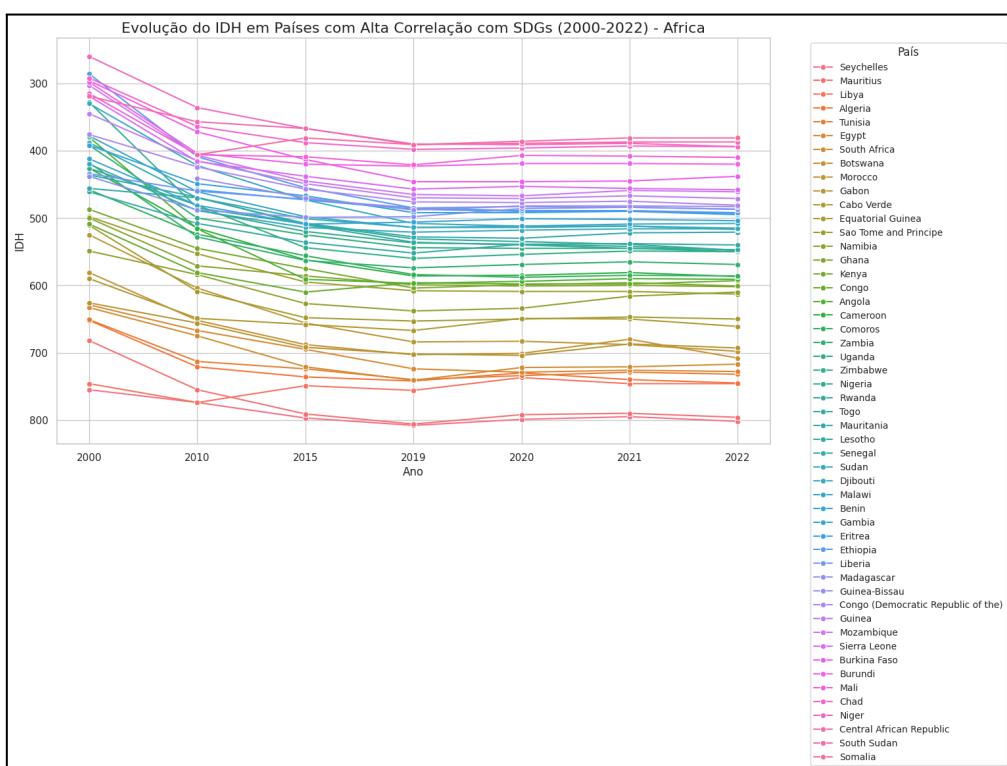
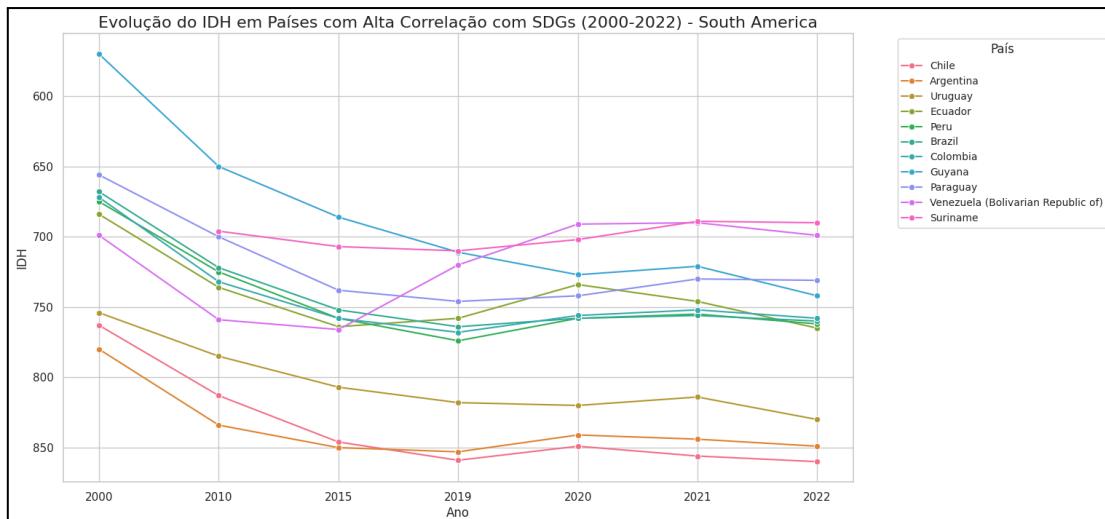
for continent in continents:
    plt.figure(figsize=(14, 8))
    sns.lineplot(data=hdi_long[hdi_long['Continent'] == continent], x='Year',
y='HDI', hue='Country', marker='o')
    plt.title(f'Evolução do IDH em Países com Alta Correlação com SDGs  
(2000-2022) - {continent}', fontsize=16)
    plt.xlabel('Ano', fontsize=12)
    plt.ylabel('IDH', fontsize=12)
    plt.legend(title='País', bbox_to_anchor=(1.05, 1), loc='upper left',
fontsize='small')
    plt.grid(True)
    plt.show()

```

Como resultado, obtivemos as seguintes representações gráficas:







Em seguida, procuramos visualizar cada país individualmente, a fim de verificar quais tiveram maior crescimento em razão dessa correlação, a fim de entender quais ações seus governos colocaram em prática para ter a melhora indicada no IDH. Assim, segue o código utilizado:

```
from tabulate import tabulate

years = ['HDI 2000', 'HDI 2010', 'HDI 2015', 'HDI 2019', 'HDI 2020', 'HDI 2021', 'HDI 2022']
sdg_columns = [
    '(SDG3) Life expectancy at birth 2022',
```

```

    '(SDG4.3) Expected years of schooling 2022',
    '(SDG4.4) Mean years of schooling 2022',
    '(SDG8.5) Gross national income per capita 2022'
]
selected_columns = years + sdg_columns

hdi_filtered = hdi[selected_columns]
correlation_matrix = hdi_filtered.corr()

correlated_columns = correlation_matrix[years].apply(lambda x: x >
0.8).any(axis=1)
correlated_columns = correlated_columns[correlated_columns].index.tolist()

correlated_data = hdi[['Country', 'Continent']] + correlated_columns

# ajuste de formato
hdi_long = pd.melt(correlated_data, id_vars=['Country', 'Continent'],
value_vars=years, var_name='Year', value_name='HDI')

hdi_long['Year'] = hdi_long['Year'].str.extract(r'(\d{4})')[0]

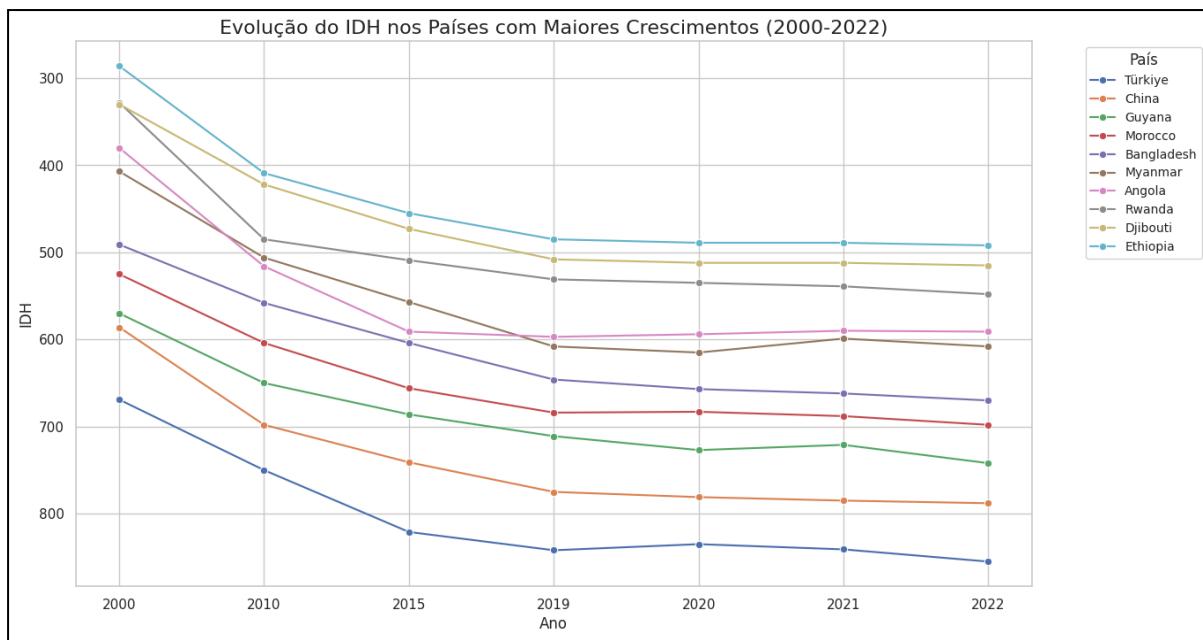
growth_data = correlated_data.set_index('Country')[years]
growth_data['Growth'] = growth_data['HDI 2022'] - growth_data['HDI 2000']
top_growth_countries = growth_data.sort_values(by='Growth',
ascending=False).head(10)

plt.figure(figsize=(14, 8))
sns.lineplot(data=hdi_long[hdi_long['Country'].isin(top_growth_countries.index)], x='Year', y='HDI', hue='Country', marker='o')
plt.title('Evolução do IDH nos Países com Maiores Crescimentos (2000-2022)', fontsize=16)
plt.xlabel('Ano', fontsize=12)
plt.ylabel('IDH', fontsize=12)
plt.legend(title='País', bbox_to_anchor=(1.05, 1), loc='upper left',
fontsize='small')
plt.grid(True)
plt.show()

```

Feito isso, verificou-se que o top 10 países com maiores crescimentos no IDH entre os anos de 2000 e 2022, quando correlacionados com os requisitos de SDG maior que 80%, foram:

- | | | |
|-------------|------------|---------------|
| 1. Rwanda | 4. China | 7. Djibouti |
| 2. Angola | 5. Myanmar | 8. Bangladesh |
| 3. Ethiopia | 6. Türkiye | 9. Morocco |
| | | 10. Guyana |



Assim, em seguida, analisamos se os primeiros 4 países, que tiveram os maiores crescimentos no IDH, implementaram políticas específicas de desenvolvimento econômico, educação ou saúde, identificando eventos históricos que possam ter impactado o IDH, como crises econômicas, guerras, ou políticas governamentais significativas.

A. Rwanda

- Desenvolvimento Pós-Conflito: Após o genocídio de 1994, o governo de Rwanda implementou várias políticas de reconstrução e desenvolvimento.
- Políticas de Saúde: Implementação de um sistema de seguro de saúde comunitário, o "Mutuelles de Santé", que aumentou o acesso aos serviços de saúde.
- Educação: Aumento do investimento em educação, com foco na educação primária e na igualdade de gênero.
- Eventos Históricos: Estabilização política e o governo de Paul Kagame focado em desenvolvimento sustentável e combate à corrupção.

B. Angola

- Pós-Guerra Civil: Fim da guerra civil em 2002, permitindo a estabilização e reconstrução do país.
- Petróleo e Economia: A indústria do petróleo impulsionou o crescimento econômico, sendo Angola um dos maiores produtores de petróleo da África.

- Investimento em Infraestrutura: Grande investimento em infraestrutura, incluindo estradas, escolas e hospitais.
- Eventos Históricos: Reconstrução pós-guerra civil e desenvolvimento do setor de petróleo.

C. Ethiopia

- Desenvolvimento Agrícola: Foco no desenvolvimento agrícola e programas de segurança alimentar.
- Políticas de Saúde: Expansão dos serviços de saúde básica e treinamento de trabalhadores de saúde comunitários.
- Educação: Melhoria no acesso à educação primária e secundária.
- Eventos Históricos: Reformas econômicas e políticas para atrair investimento estrangeiro, estabilidade política relativa.

D. China

- Reformas Econômicas: Reformas econômicas iniciadas na década de 1980 e intensificadas nos anos 2000.
- Políticas de Saúde: Expansão da cobertura de saúde pública.
- Educação: Investimentos massivos em educação e inovação tecnológica.
- Eventos Históricos: Entrada na Organização Mundial do Comércio (OMC) em 2001, crescimento econômico robusto e urbanização rápida.

A análise sugere que os países com maiores crescimentos no IDH entre 2000 e 2022 implementaram uma combinação de políticas de desenvolvimento econômico, saúde e educação. Esses países também passaram por eventos históricos significativos que catalisaram esses crescimentos. A estabilidade política, as reformas econômicas, os investimentos em infraestrutura, saúde e educação, e a integração global foram fatores chave para o aumento do IDH.