

Análise de Dados do Banco Mundial (World Bank)

Ciência dos Dados 2025.2

Francisca Maria Rodrigues Andrade - 564538

Gabriel Silva Nascimento - 571484

João Paulo Sousa Menezes - 557497

Maria Fernanda Vasconcelos Nunes - 554599

Universidade Federal do Ceará (UFC) - Campus Crateús

Novembro de 2025

[clique aqui para acessar nosso repositório](#)

1 Introdução

A América do Sul é um continente marcado por grandes contrastes. Mesmo compartilhando fronteiras, os países da região frequentemente apresentam realidades sociais, econômicas e de qualidade de vida muito distintas.

Compreender os padrões por trás dessas diferenças é o primeiro passo para entender os desafios e o potencial do continente. Este trabalho propõe uma análise exploratória focada em investigar essas dinâmicas, utilizando dados públicos e confiáveis do Banco Mundial.

O objetivo central foi construir um conjunto de dados específico para a região, cobrindo o período de 2000 a 2022. Com ele, buscamos comparar a evolução dos países e encontrar possíveis relações entre indicadores de saúde, economia e infraestrutura, oferecendo um panorama dos avanços e desigualdades sul-americanas nas últimas duas décadas.

2 Metodologia e Dados

2.1 Fonte e Coleta dos Dados

Para a construção do nosso conjunto de dados (dataset) para análise, utilizamos a biblioteca `wbgapi`, uma API que fornece acesso ao banco de dados do World Bank Open Data, e Python como linguagem de programação.

2.2 Descrição do Conjunto de Dados

O conjunto de dados é composto por 16 indicadores (atributos) selecionados, detalhados na Tabela 1. Considerando 12 países e 23 anos, o conjunto de dados brutos possui 276 instâncias (linhas) por indicador.

Tabela 1: Indicadores Selecionados para a Análise (2000-2022)

Nome Mapeado	Indicador Original (World Bank)
Acesso à Eletricidade	Access to electricity
Fertilidade Adolescente	Adolescent fertility rate
Gasto com Saúde (% PIB)	Current health expenditure
Exportação de Combustível (%)	Fuel exports
PIB per capita	GDP per capita
Expectativa de Vida	Life expectancy at birth, total
Mortalidade Infantil	Mortality rate, under-5
Acesso à Água Potável	People using at least basic drinking water
Acesso ao Saneamento Básico	People using at least basic sanitation
População Total	Population, total
Taxa de Pobreza (US\$3/dia)	Poverty headcount ratio
Prevalência de Desnutrição	Prevalence of undernourishment
Desemprego Feminino	Unemployment, female
Desemprego Masculino	Unemployment, male
Proteção Social (%)	Coverage of social protection
Inflação (% anual)	Inflation, consumer prices

2.3 Licença de Uso e Considerações Éticas

Os dados do Banco Mundial são classificados como Open Data (Dados Abertos). A maioria dos conjuntos de dados, incluindo os Indicadores de Desenvolvimento Mundial, é licenciada sob a Creative Commons Attribution 4.0. Esta licença permite que qualquer pessoa copie, distribua, adapte e use os dados, mesmo para fins comerciais, desde que a devida atribuição seja dada ao Banco Mundial.

Os dados utilizados são agregados em nível nacional e totalmente públicos. Eles não contêm Informações de identificação pessoal e não permitem a identificação de indivíduos, domicílios ou empresas específicas. A análise se restringe a indicadores macroeconômicos e sociais anonimizados, o que mitiga riscos éticos.

2.4 Pré-processamento

O pré-processamento foi realizado em Python, utilizando as bibliotecas Pandas e wbgapi. As etapas incluíram:

- Renomeação dos indicadores para nomes descritivos (ex: 'SI.POV.LMIC' para 'Taxa de Pobreza').
- Tratamento de valores ausentes (NaNs) por meio de interpolação linear, assumindo uma tendência gradual entre os anos e as diferenças entre os países.
- Reestruturação dos dados de wide para long para um formato adequado à análise de séries temporais.
- Renomeação dos indicadores para serem mencionados em português.

3 Resultados e Principais Achados

Aqui são apresentadas as principais visualizações geradas a partir dos dados tratados.

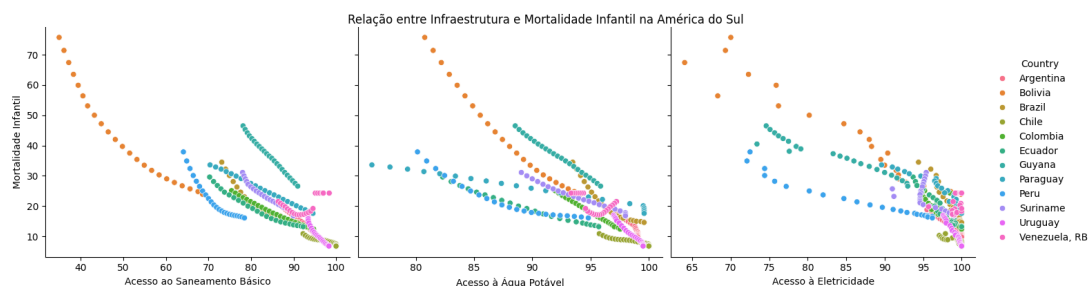


Figura 1: Evidência da relação entre o desenvolvimento de infraestrutura básica (saneamento, água potável, eletricidade) e a taxa de mortalidade infantil.

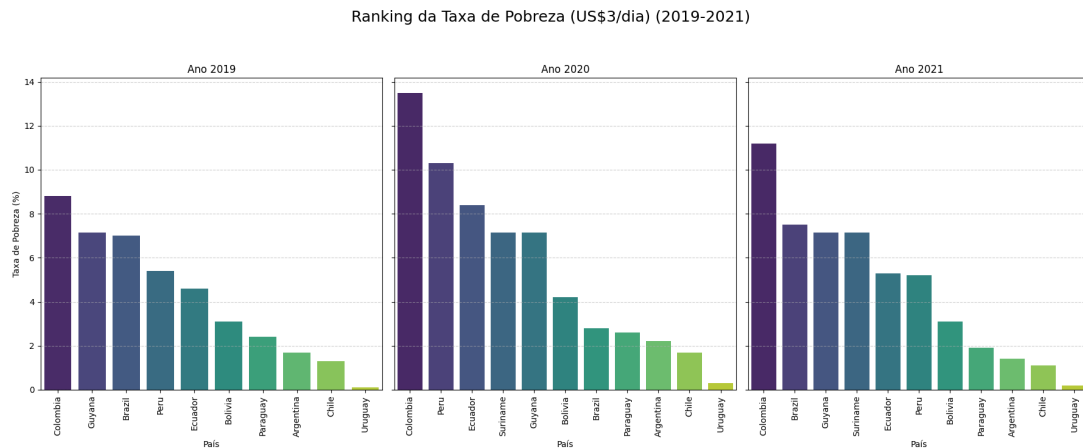


Figura 2: Representação das taxas de pobreza dos países da América do Sul no período 2019-2021, incluindo períodos pré, durante e pós pandemia. Mostrando a grande instabilidade da posição do Brasil.

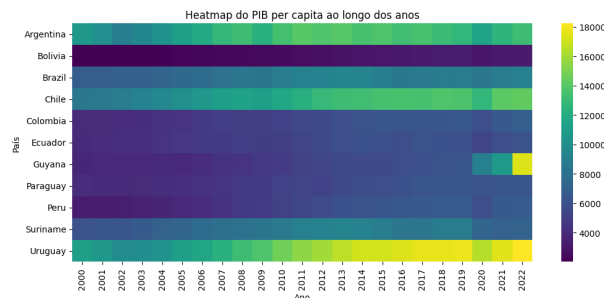


Figura 3: Representação por mapa de calor que evidencia um aumento expressivo no PIB da Guyana.

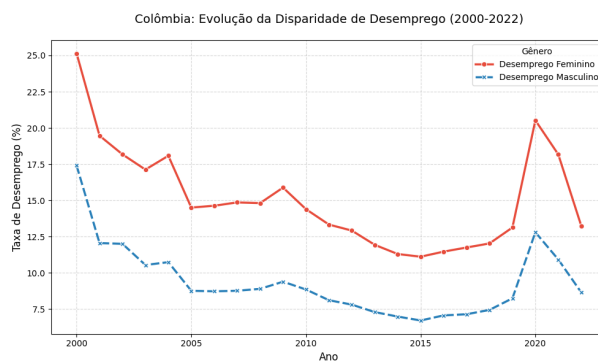


Figura 4: Acentuada discrepância entre as taxas de desemprego feminina e masculina na Colômbia durante o ano de 2020.

4 Desafios e Aprendizados

4.1 Principais Desafios

Durante o desenvolvimento do projeto, enfrentamos alguns obstáculos técnicos e metodológicos que foram importantes para o aprendizado:

- **Tratamento de Dados Ausentes (NaNs):** O primeiro desafio foi técnico e estatístico. A base de dados do Banco Mundial apresentou falhas (valores ‘NaN’), como qualquer outra base de dados. Nosso desafio foi escolher a melhor forma de imputação desses dados ausentes.
- **Gestão da Consistência dos Dados (Pipeline):** Manter a consistência dos nomes das variáveis (indicadores) e dos recursos visuais ao longo de todo o notebook foi um desafio de programação, haja vista a quantidade de pessoas utilizando e editando.
- **Visualização Eficaz (Storytelling):** Além da dificuldade técnica de implementar certos gráficos (como os de eixo duplo ou facetados), o maior desafio foi metodológico. Houve uma dificuldade em traduzir os dados brutos em visualizações que contassem uma história clara e gerassem insights notáveis.

4.2 Principais Aprendizados

Os desafios enfrentados se converteram em aprendizados práticos para o grupo:

- **Consumo de APIs e Pré-processamento:** Uso prático da biblioteca ‘wbgapi’ para coleta de dados. Além de tratamento de dados e análise exploratória inicial.
- **Visualização de Dados:** Criação de gráficos com vários parâmetros e informações e utilização de bibliotecas de criação visual de dados.
- **Boas Práticas em Análise de dados:** Aprendemos a importância da padronização de variáveis e entender o que as outras pessoas da equipe fizeram.

4.3 Limitações

As principais limitações deste trabalho foram de natureza técnica e metodológica:

- **Imprecisão da Interpolação:** A estratégia usada para lidar com os dados ausentes, a interpolação linear, foi uma de nossas limitações. Este método assume um crescimento constante entre os pontos de dados e não captura a volatilidade real de crises, por exemplo.
- **Correlação não Implica Causalidade:** Este estudo teve um caráter exploratório e visual. Os gráficos apresentados podem sugerir fortes correlações entre indicadores, mas não estabelecem, necessariamente, causalidade estatística.

4.4 Recomendações para Trabalhos Futuros

Com base nas limitações identificadas, sugerem-se as seguintes direções para trabalhos futuros:

- **Enriquecimento do Dataset:** Para resolver o problema dos dados ausentes, sugere-se o cruzamento dos dados do Banco Mundial com outras fontes.
- **Desenvolvimento de Visualizações Interativas:** No lugar de gráficos estáticos (.png), um trabalho futuro poderia focar na criação de visualizações interativas.