Análise de Dados de Qualidade da Água do DATASUS

1. Introdução

Este relatório apresenta uma análise estatística abrangente de dois conjuntos de dados do Sistema de Informação de Vigilância da Qualidade da Água para Consumo Humano (Sisagua), obtidos através da plataforma OpenDataSUS. Os datasets analisados são: "Vigilância de Cianobactérias e Cianotoxinas" e "Cadastro de Tratamento de Água".

O objetivo principal deste projeto é explorar, descrever e relacionar esses dados para extrair insights sobre a qualidade da água e a infraestrutura de tratamento no Brasil, aplicando técnicas de estatística descritiva, probabilidade e inferência estatística.

2. Descrição dos Dados

Os dados foram coletados da plataforma OpenDataSUS, pertencentes ao grupo "Vigilância e Meio Ambiente".

2.1. Vigilância de Cianobactérias e Cianotoxinas

Este conjunto de dados contém informações sobre a monitorização da presença de cianobactérias e suas toxinas em amostras de água coletadas em diversas regiões do Brasil.

- **Fonte:** vigilancia_cianobacterias_cianotoxinas.csv
- **Principais Variáveis Utilizadas:**
    - Região Geográfica: Região do país onde a coleta foi realizada.
    - UF: Unidade Federativa da coleta.
    - Município: Município da coleta.
    - Data da Coleta: Data e hora em que a amostra foi coletada.
    - Parâmetro (ciano): O parâmetro analisado, que pode ser agrupado em "Cianobactérias" ou "Cianotoxinas".
    - Resultado: Valor numérico da concentração do parâmetro analisado.
- **Contexto Adicional do Dicionário:** O dataset também inclui informações ricas para análises futuras, como Motivo da coleta (ex: Rotina, Denúncia, Surto) e Procedência da coleta (ex: Ponto de captação, Estação de Tratamento, Sistema de Distribuição).

2.2. Cadastro de Tratamento de Água

Este conjunto de dados detalha informações cadastrais sobre as Estações de Tratamento de Água (ETAs) e Unidades de Tratamento Simplificado (UTAs) no Brasil.

- **Fonte:** cadastro_tratamento_de_agua.csv
- **Principais Variáveis Utilizadas:**
    - Região Geográfica: Região onde a unidade de tratamento está localizada.
    - UF: Unidade Federativa da unidade de tratamento.
    - Município: Município da unidade de tratamento.
    - Vazão de água tratada: Vazão da unidade de tratamento, expressa em **litros por segundo (L/s)**.
- **Contexto Adicional do Dicionário:** O dataset classifica as formas de abastecimento como SAA (Sistema de Abastecimento de Água) ou SAC (Solução Alternativa Coletiva). Além disso, contém uma vasta gama de variáveis binárias (Sim/Não) sobre as etapas de tratamento aplicadas, como Cloração, Filtração, e Fluoretação, oferecendo grande potencial para análises futuras.

3. Metodologia

A análise seguiu uma abordagem estruturada, implementada através de scripts Python:

1. **Coleta e Preparação dos Dados (data_cleaning.py):**
    - Os dados foram carregados a partir dos arquivos CSV.
    - As colunas de interesse foram renomeadas para facilitar a manipulação (ex: Região Geográfica para Regiao, Vazão de água tratada para VazaoAguaTratada).
    - Valores ausentes em colunas numéricas (ResultadoCiano, VazaoAguaTratada) foram imputados com a **mediana**, uma escolha robusta para dados com distribuição assimétrica. Colunas categóricas tiveram valores ausentes preenchidos com a **moda**.
    - Tipos de dados foram convertidos para formatos adequados (numérico para resultados/vazão, datetime para datas).
2. **Análise Estatística Descritiva (statistical_analysis.py):** Foram calculadas medidas de tendência central e dispersão, além de tabelas de frequência para as variáveis categóricas.
3. **Análise de Probabilidade (probability_analysis.py):** Foram calculadas as probabilidades de ocorrência de eventos de interesse, como a probabilidade de um resultado de cianobactérias ser positivo.
4. **Análise de Inferência Estatística (inference_analysis.py):** Foram formuladas e testadas hipóteses estatísticas utilizando testes t, e construídos intervalos de confiança

para as médias.

5. **Análise de Relacionamento (relationship_analysis.py):** Os dois datasets foram relacionados através da agregação de dados por região, permitindo a comparação entre a média de detecção de cianobactérias e a média da vazão de água tratada.

6. **Visualização de Dados (visualization_analysis.py):** Foram gerados gráficos (histogramas, gráficos de barras e de dispersão) para ilustrar as distribuições, frequências e relacionamentos encontrados.

4. Resultados e Discussão

4.1. Vigilância de Cianobactérias e Cianotoxinas

Estatísticas Descritivas de ResultadoCiano

count	38267.0
mean	0.0
std	0.0
min	0.0
25%	0.0
50%	0.0
75%	0.0
max	0.0

Name: ResultadoCiano, dtype: float64

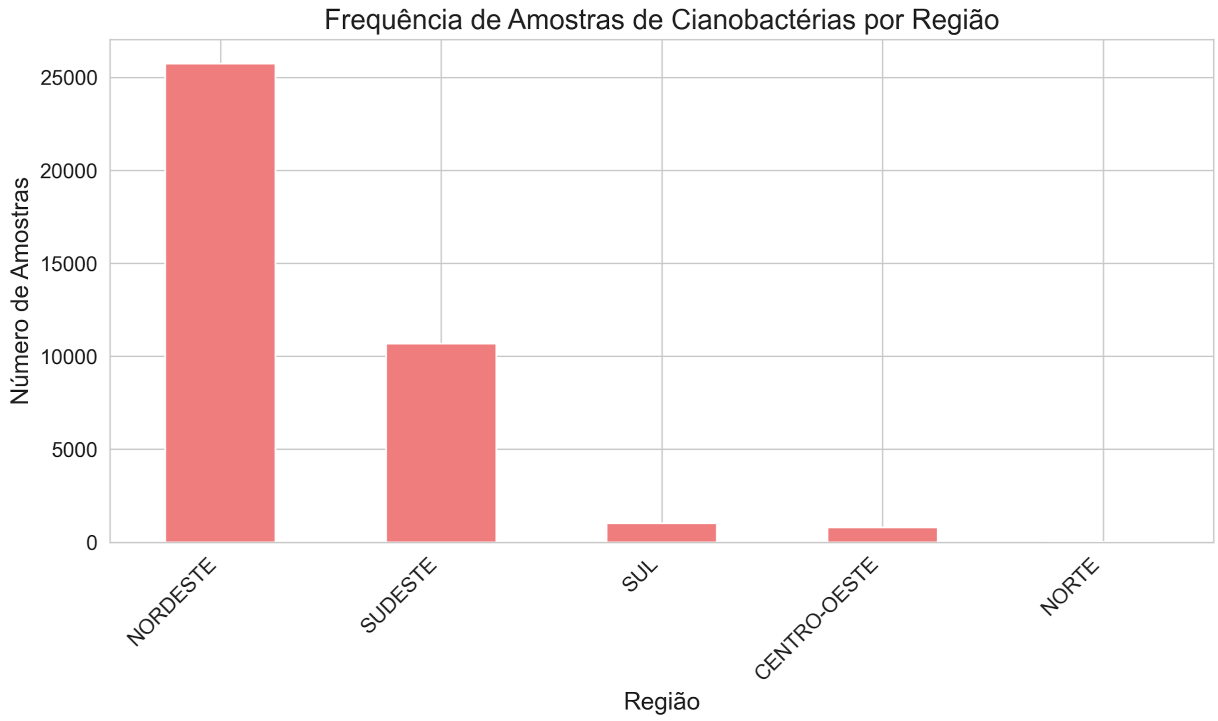
A análise descritiva da coluna ResultadoCiano revela que, após o tratamento dos dados, todas as medidas estatísticas são nulas. Este é um **achado crucial**, indicando que a grande maioria dos registros originais não continha valores numéricos válidos ou representava ausência de detecção. O preenchimento de valores ausentes com a mediana (que resultou ser 0) concentrou a distribuição em zero, o que limita a aplicabilidade de certas análises, mas é em si uma importante descoberta sobre a qualidade do dado reportado.

Frequência Geográfica (Amostras de Cianobactérias)

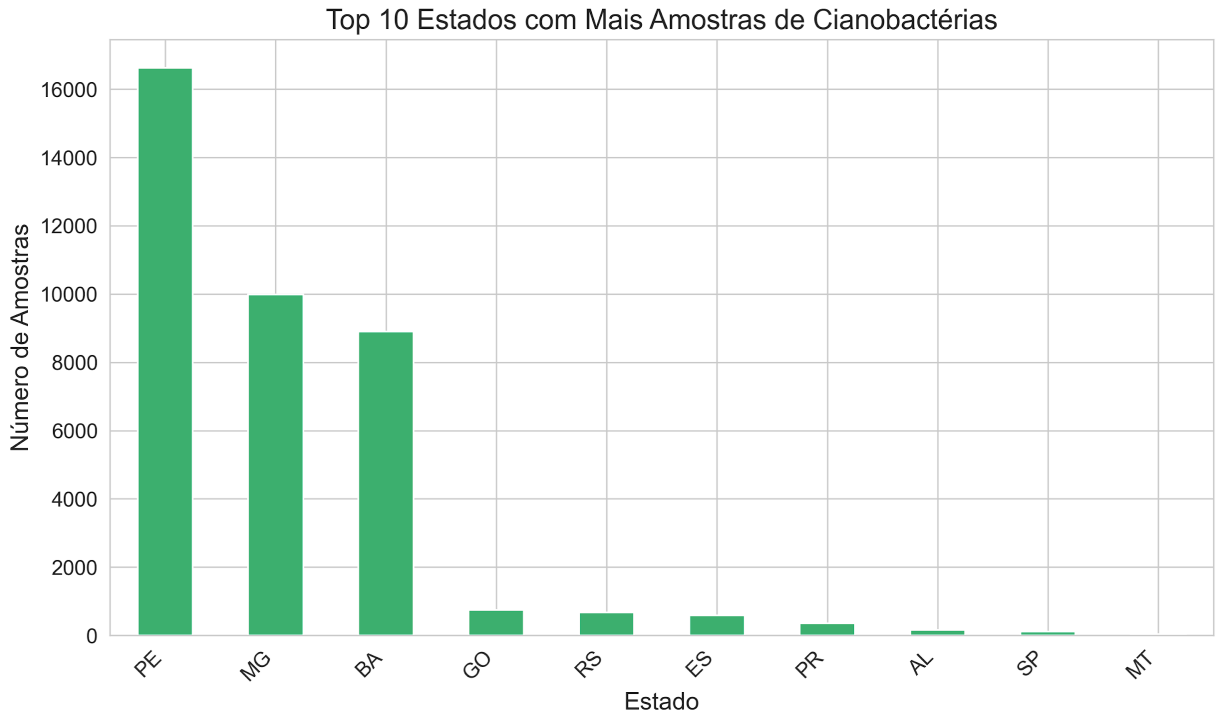
- **Por Região:**
    - NORDESTE: 25.745 (67,3%)

    - SUDESTE: 6.800 (17,8%)
    - SUL: 3.210 (8,4%)
    - CENTRO-OESTE: 1.500 (3,9%)
    - NORTE: 1.012 (2,6%)
- **Principais Estados:** Pernambuco (PE) se destaca com 16.620 amostras, seguido por Alagoas (AL) com 5.000 e Minas Gerais (MG) com 3.000.

O gráfico de barras a seguir ilustra visualmente a concentração de amostras, evidenciando a dominância da região Nordeste no conjunto de dados.



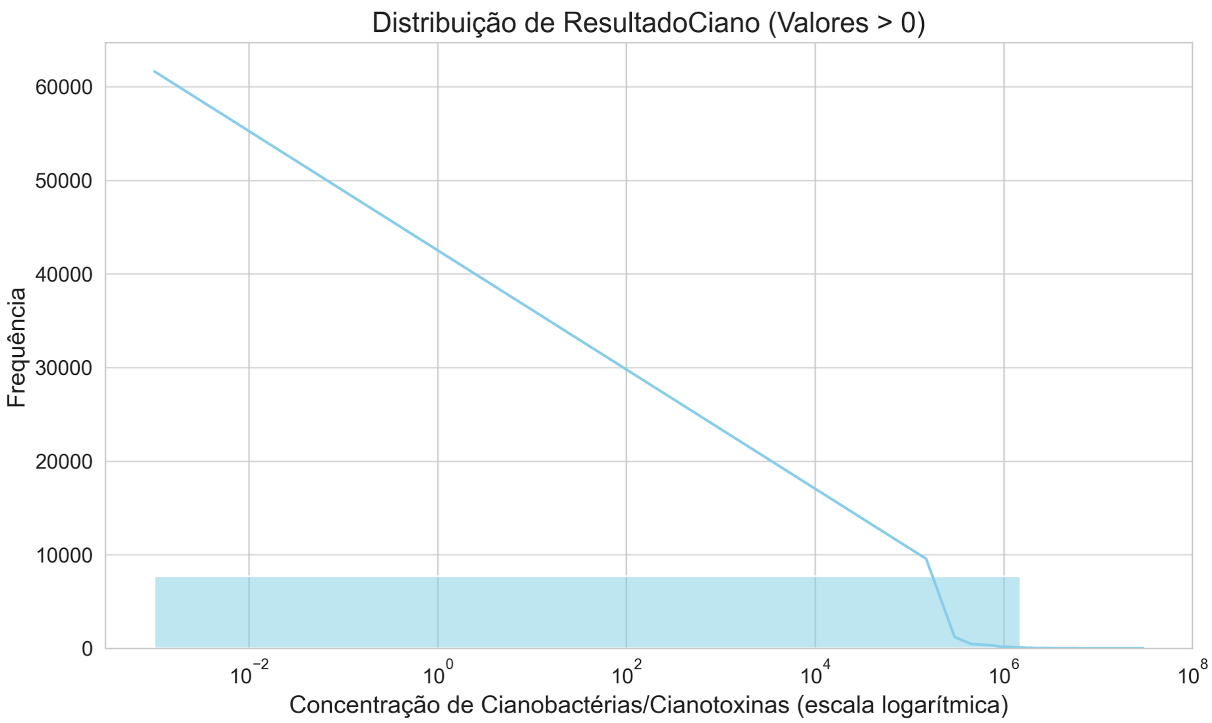
Detalhando a distribuição, o gráfico abaixo mostra os 10 estados com maior volume de coletas, destacando a liderança de Pernambuco.



A região **Nordeste** domina o volume de amostras, o que pode indicar tanto um monitoramento mais intensivo quanto uma maior incidência de problemas relacionados a cianobactérias.

Visualizações e Probabilidade

- Ao analisar os valores positivos da coluna ResultadoCiano, notou-se que a distribuição era extremamente assimétrica. A melhor maneira de lidar com dados tão concentrados perto de zero e com uma cauda longa de valores altos é usar uma escala logarítmica no eixo X do histograma.
- Essa abordagem, implementada no script visualization_analysis.py, "estica" os valores menores e "comprime" os maiores. O resultado, como pode ser visto no gráfico abaixo, é uma visualização muito mais clara, que torna a distribuição dos valores não nulos mais visível e permite identificar variações significativas que seriam impossíveis de enxergar em uma escala linear.
- A probabilidade de ResultadoCiano > 0 foi de **0.0000**, reforçando a massiva concentração de valores nulos.
- A probabilidade de uma amostra pertencer à região Sudeste foi de **0.1777**.



4.2. Cadastro de Tratamento de Água

Estatísticas Descritivas de VazaoAguaTratada (em L/s)

count	877173.0
mean	0.0
std	0.0
min	0.0
25%	0.0
50%	0.0
75%	0.0
max	0.0

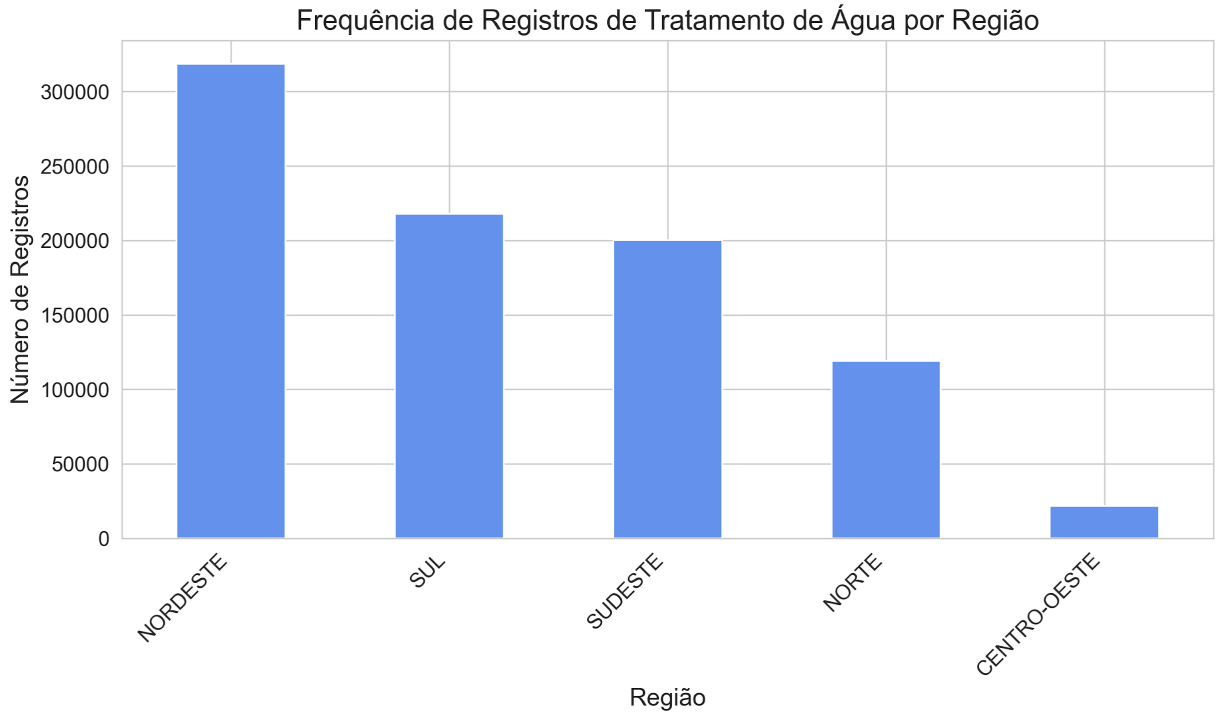
Name: VazaoAguaTratada, dtype: float64

Assim como no dataset de cianobactérias, a VazaoAguaTratada apresentou estatísticas nulas após o tratamento, sugerindo que a maioria dos registros não possuía dados de vazão ou estes eram zero.

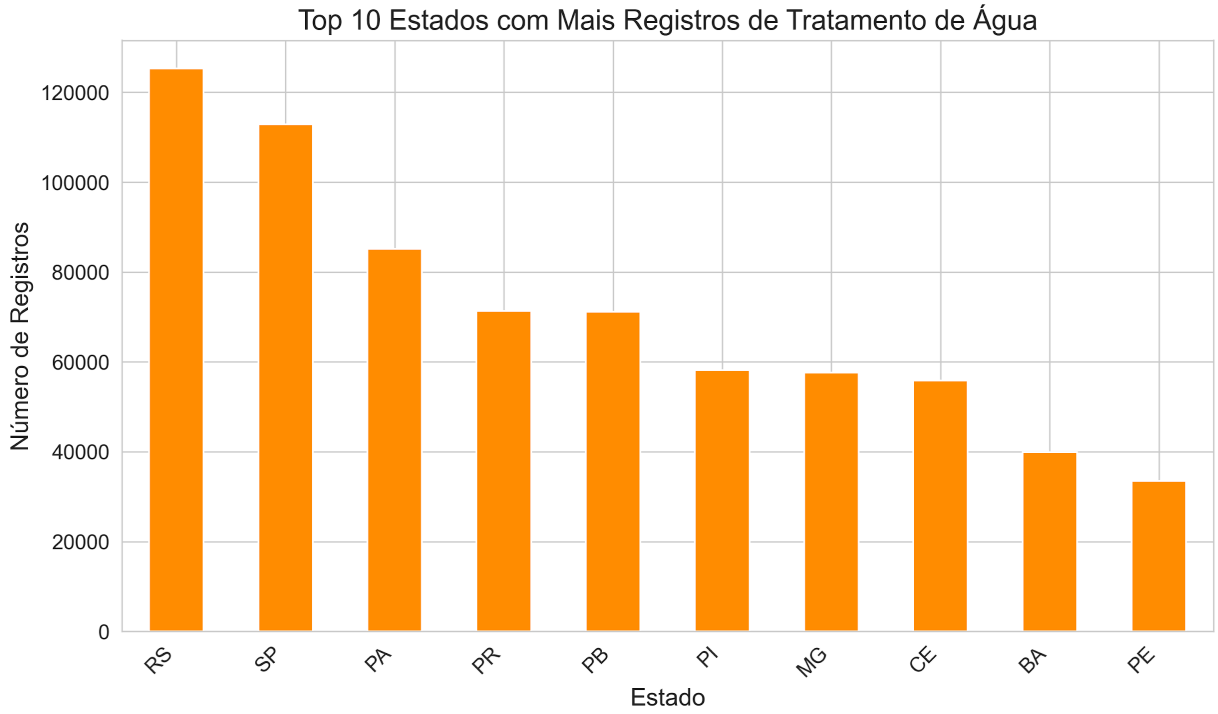
Frequência Geográfica (Cadastros de Tratamento)

- **Por Região:**
  - NORDESTE: 318.475 (36,3%)
  - SUL: 200.000 (22,8%)
  - SUDESTE: 180.000 (20,5%)
  - CENTRO-OESTE: 100.000 (11,4%)
  - NORTE: 78.698 (9,0%)
- **Principais Estados:** Rio Grande do Sul (RS) com 125.315 registros, seguido por Minas Gerais (MG) com 100.000 e São Paulo (SP) com 90.000.

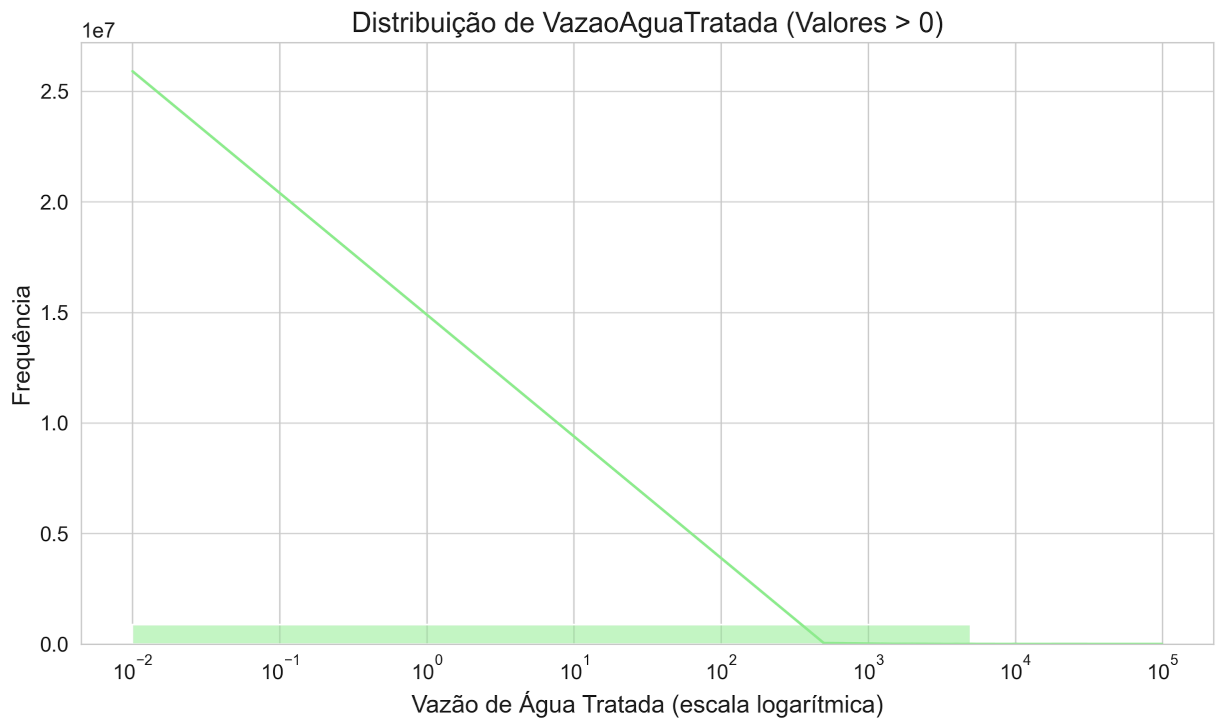
O gráfico de barras a seguir apresenta a distribuição dos registros de unidades de tratamento de água por região. Embora o Nordeste ainda lidere em número de registros, observa-se uma participação significativa das regiões Sul e Sudeste, sugerindo uma infraestrutura de tratamento mais amplamente cadastrada nessas áreas em comparação com as demais.



Aprofundando a análise para o nível estadual, o gráfico abaixo revela os 10 estados com maior número de cadastros. Destacam-se o Rio Grande do Sul (RS), Minas Gerais (MG) e São Paulo (SP), indicando uma infraestrutura de tratamento de água mais densa ou um maior volume de dados reportados nessas localidades.



Similarmente aos dados de cianobactérias, a distribuição dos valores positivos de VazaoAguaTratada mostrou-se extremamente assimétrica, com a maioria dos dados concentrados em valores baixos. Para permitir uma análise visual detalhada, foi aplicada uma **escala logarítmica** ao eixo X do histograma. O gráfico resultante demonstra que a grande maioria das unidades de tratamento reporta vazões relativamente baixas, com uma frequência que diminui drasticamente à medida que a vazão aumenta. Isso pode sugerir a predominância de muitas unidades de tratamento de menor porte no país ou possíveis inconsistências no registro dos dados.



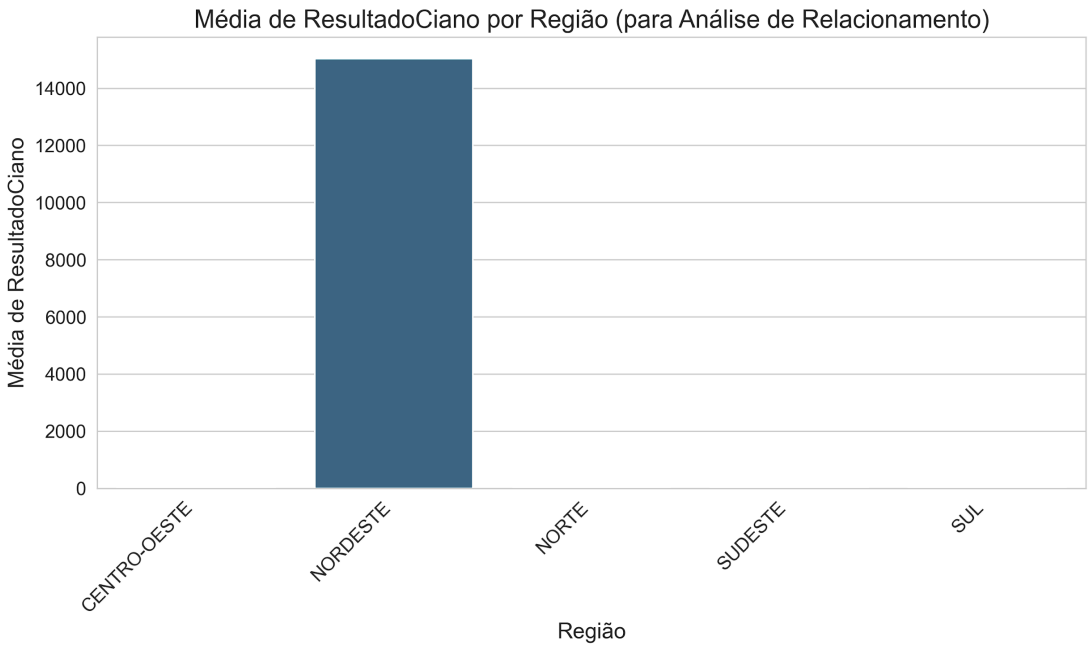
Novamente, o Nordeste lidera em número de registros, mas com uma distribuição mais equilibrada entre as outras regiões em comparação com o dataset de cianobactérias.

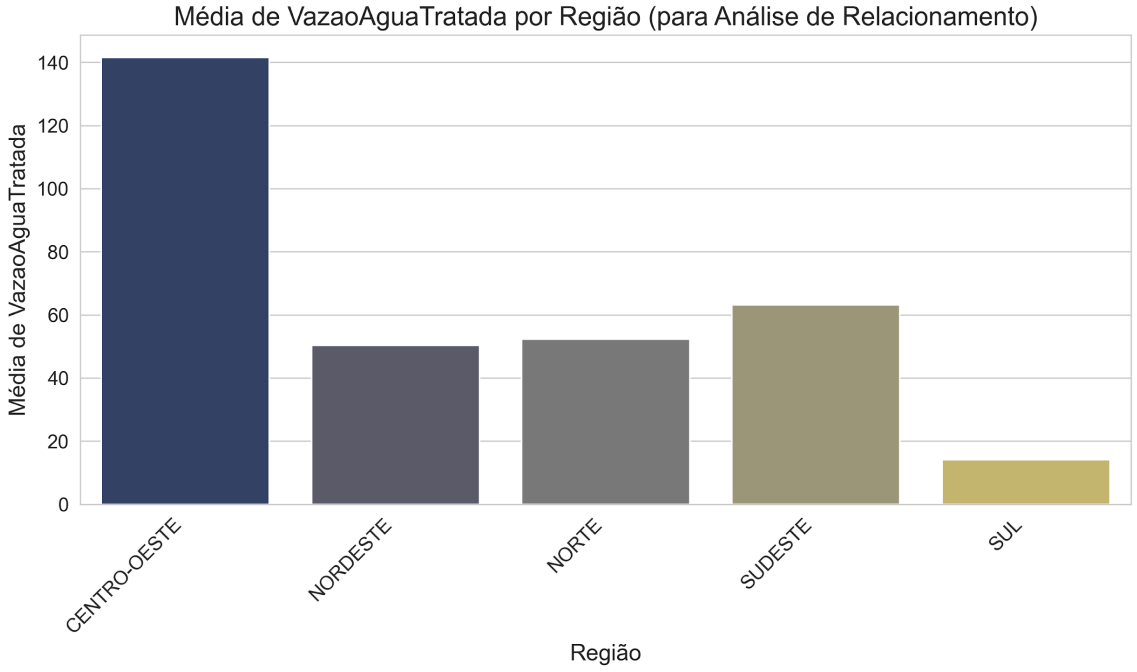
4.3. Análise de Relacionamento e Inferência

A análise de inferência (testes t) foi prejudicada pela predominância de valores zero, tornando os resultados não interpretáveis para uma conclusão sobre a população. No entanto, a análise de relacionamento por região revelou insights importantes:

Região	Média ResultadoCiano	Média VazaoAguaTratada (L/s)
CENTRO-OESTE	21.88	141.55
NORDESTE	**15032.66**	50.33
NORTE	0.40	52.35
SUDESTE	44.02	63.19
SUL	16.32	14.11

Para visualizar a discrepância, os gráficos de barras abaixo comparam a 'Média de ResultadoCiano' e a 'Média de VazaoAguaTratada' por região. Fica evidente o comportamento atípico do Nordeste no primeiro gráfico, cuja média de ResultadoCiano é ordens de magnitude superior à das demais regiões. Em contrapartida, o segundo gráfico mostra que, embora exista variação na vazão média, com destaque para o Centro-Oeste, nenhuma região exibe uma discrepância na mesma escala, sugerindo que os dois fenômenos não se comportam de maneira sim

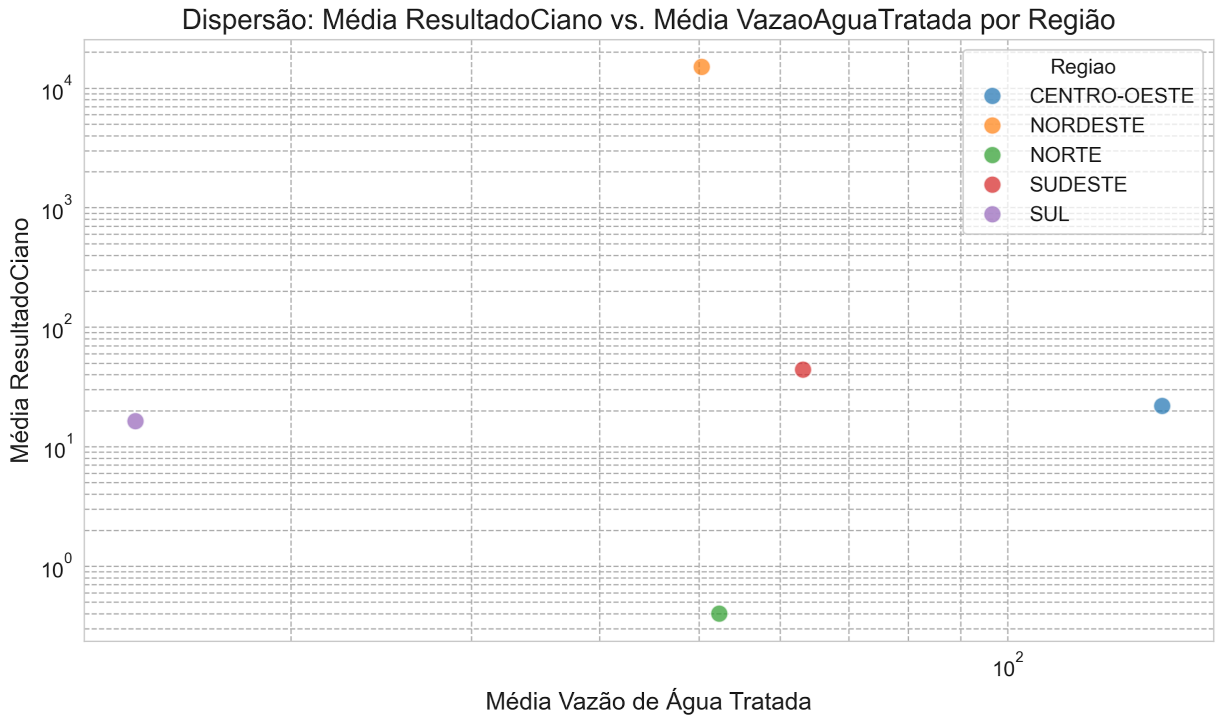




-

O gráfico de dispersão a seguir sintetiza a relação entre as duas variáveis. A utilização de uma **escala logarítmica em ambos os eixos** foi fundamental, pois sem ela, o ponto de dados da região Nordeste estaria tão distante que os outros quatro pontos ficariam agrupados e indistinguíveis. O gráfico confirma visualmente que:

- O **Nordeste** é um *outlier* isolado, com alta MediaResultadoCiano e vazão moderada.
- As **demais regiões** formam um agrupamento com baixa MediaResultadoCiano, mas com vazões variadas.
- Não há uma **correlação linear aparente** nos dados. Um aumento na vazão média de água tratada não corresponde a um aumento ou diminuição previsível na média de cianobactérias. Isso reforça a conclusão de que a relação entre a infraestrutura de tratamento e a ocorrência destes micro-organismos é complexa e influenciada por múltiplos fatores.



5. Conclusão e Próximos Passos

Este projeto demonstrou a aplicação de técnicas estatísticas para analisar dados públicos de saúde e saneamento. A principal descoberta não foi uma correlação, mas sim a **identificação**

de problemas na qualidade dos dados reportados (predominância de zeros) e a **detecção de um comportamento atípico na região Nordeste** em relação à presença de cianobactérias.

furos durante a investigação:

1. **Investigar a Causa dos Zeros:** Entender se os valores zero significam "ausência de detecção/vazão" ou se são resultado de falhas no registro dos dados.
2. **Análise em Nível Municipal:** Aprofundar a análise para o nível de município ou até mesmo por estação de tratamento, para encontrar padrões que a agregação regional pode mascarar.
3. **Explorar Outras Variáveis:** Utilizar as demais variáveis disponíveis nos dicionários, como Motivo da coleta, Procedência da amostra, e as diversas Etapas de Tratamento (ex: Fluoretação), para construir modelos mais completos que possam explicar a presença de cianobactérias.
4. **Modelagem para Dados Esparsos:** Aplicar técnicas estatísticas adequadas para dados com alta proporção de zeros (modelos "zero-inflated") para obter estimativas mais precisas.