Previsão Mensal de Emissão de Dióxido de Carbono na Amazônia Legal

**Luísa Ferreira da Silveira (2210875)**

**Pedro de Almeida Barizon (2211350)**

**Theo Couto Xavier (2210610)**

**Vinícius Lucena Bitu Cortez (2210458)**

**Professora:** Paula Medina Maçaira Louro

**Monitora:** Laura Nascimento Coutinho

**Turma:** 33A

**Data:** 06/06/2025

# Introdução

## Motivação

Considerando o macrotema do projeto AMazonizAR, um esforço de inserção da PUC-Rio no contexto da Amazônia [ECOA PUC-RIO 2023], e a premente situação sócio-climático-ambiental que não apenas a Grande Floresta, mas todo o planeta enfrenta, optou-se pelo seguinte tema para o projeto: ***Na Amazônia Legal, qual a inter-relação entre incêndios florestais, mudanças climáticas e emissão de dióxido de carbono (CO₂)?***

## Questão de Pesquisa

A fim de transformar o tema do projeto em algo mais tangível e mensurável, formulou-se esta questão de pesquisa (QP):

***A partir dos indicadores ambientais de um estado da Amazônia Legal e seus níveis de incêndios florestais (queimadas) em um dado mês, qual a quantidade de CO₂, em toneladas, emitida por ele na atmosfera nesse mês?***

Para o caso de a absorção ser maior que a emissão, convencionou-se uma emissão negativa. Quanto aos estados, serão considerados segundo a classificação do IBGE [2020]:

* Acre (AC);
* Amapá (AP);
* Amazonas (AM);
* Maranhão (MA);
* Mato Grosso (MT);
* Pará (PA);
* Rondônia (RO);
* Roraima (RR);
* Tocantins (TO).

## Cliente a Ser Atendido

Os clientes-alvo serão os **participantes do AMazonizAR, da PUC-Rio, e eventuais órgãos ambientais e governamentais que desejem fazer uso do modelo de dados gerado**. Espera-se oferecer-lhes uma ferramenta que agregue valor ao permitir estimar um panorama de emissões da Floresta, o que configura um indicador útil em relação ao controle da crise climática e ao mundo dos créditos de carbono — temas de interesse de toda a comunidade global, que poderá mostrar-se mais solícita na realização de ações interventivas diante de estimativas quantitativas.

## Objetivo

No que diz respeito aos objetivos deste trabalho, mesclam tanto interesses de natureza acadêmica quanto de natureza cidadã. Se, por um lado, almeja-se aprender os conceitos e as ferramentas que compõem a Ciência de Dados contemporânea; por outro, objetiva-se compreender melhor o cenário amazônico, tão relevante não apenas à nação brasileira, senão ao mundo inteiro, que vê chegarem os efeitos de uma crise ambiental que não parece ter planos de ir embora. Assim, na impossibilidade de vê-la partir, deve-se aprender a conviver com a iminente crise — e, para tanto, são necessários análise e estudo.

## Metodologia a Ser Empregada

(Descrever brevemente a metodologia que será utilizada para abordar a questão de pesquisa e alcançar os objetivos do projeto. Isso pode incluir o tipo de análise de dados, as técnicas de aprendizado de máquina a serem aplicadas, entre outros.)

Para garantir a reprodutibilidade e confiabilidade do presente trabalho, foram coletadas bases de dados públicas e abertas, conforme detalhado na seção de Obtenção e Tratamento dos Dados, contendo informações sobre queimadas e variáveis climáticas nos estados da Amazônia Legal com granularidade mensal.

Em seguida, essas informações passaram por um processo de pré-processamento e enriquecimento, de modo a unificá-las em uma única base de dados consolidada. Esse tratamento envolveu desde a padronização de formatos até a imputação e correção de valores faltantes, com atenção à preservação das características temporais e espaciais relevantes para o estudo.

Com o conjunto final obtido, realizou-se uma análise exploratória (descrita na seção de Análise Exploratória dos Dados) para compreender a estrutura do dataset, investigar padrões, identificar possíveis outliers e avaliar correlações entre as variáveis envolvidas.

A etapa seguinte consistiu na aplicação de dois modelos de aprendizado de máquina supervisionado voltados à tarefa de regressão:

* **Regressão Linear Múltipla**: técnica estatística que modela a relação entre uma variável dependente contínua (neste caso, a emissão mensal de CO₂) e múltiplas variáveis independentes. Assume-se que essa relação é linear, ou seja, pode ser representada por uma combinação linear dos preditores.
* **Árvore de Regressão**: modelo baseado em uma estrutura hierárquica de decisões, em que os dados são divididos em subconjuntos progressivamente mais homogêneos com relação à variável alvo. Tal técnica é não linear e apresenta boa capacidade de capturar interações e padrões complexos entre os atributos.

Ambos os modelos foram avaliados com base nas seguintes métricas:

* Erro Quadrático Médio (MSE): métrica que penaliza fortemente erros grandes, por elevar ao quadrado a diferença entre valor real e previsto. É naturalmente minimizada pela regressão linear. A fórmula é:

MSE=1n∑i=1n(yi−y^i)2\text{MSE} = \frac{1}{n} \sum\_{i=1}^{n} (y\_i - \hat{y}\_i)^2MSE=n1​i=1∑n​(yi​−y^​i​)2

* Erro Absoluto Médio (MAE): calcula a média dos módulos dos erros. Por ser uma métrica de ordem 1, é mais robusta a outliers e preserva a unidade da variável predita, permitindo interpretação direta do erro médio. Sua fórmula é:

MAE=1n∑i=1n∣yi−y^i∣\text{MAE} = \frac{1}{n} \sum\_{i=1}^{n} |y\_i - \hat{y}\_i|MAE=n1​i=1∑n​∣yi​−y^​i​∣

* Coeficiente de Determinação (R²): quantifica a proporção da variância da variável-alvo que é explicada pelo modelo. É interpretável em uma ampla gama de contextos, e por isso amplamente adotado. Importante notar que, embora usualmente varie entre 0 e 1, o R² pode assumir valores negativos — o que indica que o modelo performa pior do que simplesmente predizer a média de yyy. Sua fórmula é:

R2=1−∑i=1n(yi−y^i)2∑i=1n(yi−yˉ)2R^2 = 1 - \frac{\sum\_{i=1}^{n} (y\_i - \hat{y}\_i)^2}{\sum\_{i=1}^{n} (y\_i - \bar{y})^2}R2=1−∑i=1n​(yi​−yˉ​)2∑i=1n​(yi​−y^​i​)2​

O uso combinado dessas métricas permitiu uma avaliação equilibrada e comparável dos modelos propostos. O MSE foi escolhido por sua coerência com o método dos mínimos quadrados, que fundamenta a regressão linear; o MAE foi incluído por sua robustez a valores extremos, característica importante diante da possível ocorrência de outliers no contexto ambiental; e o R² foi utilizado por sua interpretabilidade relativamente universal, além de ser independente da escala dos dados.

Por fim, os resultados obtidos foram analisados criticamente, com o objetivo de avaliar se os dados disponíveis permitiriam responder de forma satisfatória à questão de pesquisa. Essa avaliação foi conduzida com base no desempenho dos modelos e na análise das variáveis mais relevantes, buscando verificar em que medida os fatores ambientais e de queimadas influenciam as emissões mensais de CO₂ na Amazônia Legal.

# Metodologia e Plano de Experimentação

## Tipo de Aprendizado de Máquina e Especificação da Tarefa

(Discutir o tipo de aprendizado de máquina escolhido, como supervisionado, não supervisionado ou semi-supervisionado. Especificar a tarefa de aprendizado realizada, como classificação, regressão, clustering, entre outros, e justificar por que essa tarefa é adequada para o problema em questão.)

Dada a natureza da questão de pesquisa — que busca estimar, com base em fatores climáticos e de queimadas, a quantidade mensal de dióxido de carbono (CO₂) emitida na atmosfera por cada estado da Amazônia Legal — optou-se pelo uso de **aprendizado de máquina supervisionado**, cuja característica principal é o uso de dados rotulados para treinar os modelos.

Mais especificamente, a tarefa em questão é de **regressão**, uma vez que a variável-alvo, car\_c02\_emitido, é uma variável numérica contínua expressa em toneladas. O objetivo central é prever valores reais dessa variável a partir de um conjunto de atributos independentes também quantitativos ou transformados numericamente.

A escolha por um modelo supervisionado de regressão se impôs naturalmente, dado que se teve acesso a séries históricas com emissões já conhecidas para os estados da Amazônia Legal. Assim, havia à disposição pares de entrada e saída — isto é, registros com os fatores ambientais e de queimadas associados aos respectivos valores de emissão de CO₂ —, o que permitiu o treinamento e avaliação de modelos preditivos baseados em dados reais.

Essa abordagem mostrou-se coerente com os objetivos do projeto, que exigem não apenas a compreensão das variáveis envolvidas, mas a capacidade de prever valores futuros ou não observados com base em padrões históricos. Além disso, a regressão é a técnica estatística e computacional adequada quando se deseja estimar uma variável dependente contínua a partir de um conjunto de preditores.

## Técnicas/Algoritmos Utilizados

(Detalhar as técnicas e algoritmos de aprendizado de máquina escolhidos para resolver o problema. Descrever brevemente cada técnica/algoritmo, incluindo sua aplicação e justificativa para sua escolha.)

Para abordar o problema de predição da emissão mensal de CO₂ na Amazônia Legal com base em fatores ambientais e de queimadas, foram aplicadas duas técnicas clássicas de regressão supervisionada: a **Regressão Linear Múltipla** e a **Árvore de Decisão para Regressão**. A escolha por essas abordagens foi motivada pela combinação de três critérios principais: adequação técnica ao problema (regressão contínua), interpretabilidade dos modelos e familiaridade do grupo com os algoritmos, o que facilita sua correta implementação e análise crítica.

#### Regressão Linear Múltipla

A Regressão Linear Múltipla é uma técnica estatística que modela a relação entre uma variável dependente contínua yyy (neste caso, a emissão de CO₂) e múltiplas variáveis independentes x1,x2,…,xpx\_1, x\_2, \ldots, x\_px1​,x2​,…,xp​, assumindo uma relação linear entre elas. O modelo é expresso pela equação:

y=β0+β1x1+β2x2+⋯+βpxp+εy = \beta\_0 + \beta\_1 x\_1 + \beta\_2 x\_2 + \cdots + \beta\_p x\_p + \varepsilony=β0​+β1​x1​+β2​x2​+⋯+βp​xp​+ε

Nessa equação, β0\beta\_0β0​ representa o intercepto, βi\beta\_iβi​ os coeficientes de regressão que indicam o impacto de cada variável preditora na variável-alvo, e ε\varepsilonε o erro aleatório. A estimação dos coeficientes se dá, usualmente, pelo método dos mínimos quadrados, que minimiza a soma dos quadrados dos resíduos (diferença entre valores reais e previstos).

Essa técnica foi escolhida por sua **simplicidade e interpretabilidade**, o que facilita a análise dos efeitos individuais das variáveis climáticas e de queimadas sobre as emissões de CO₂. Além disso, tem como aplicação permitir testar a significância estatística de cada variável — dada por seu respectivo coeficiente beta\_i —, ajudando na seleção de atributos relevantes.

#### Árvore de Regressão (Decision Tree Regressor)

A Árvore de Regressão é um algoritmo baseado em uma estrutura de decisão hierárquica, em que o espaço dos dados é particionado de forma recursiva com base em divisões que minimizam o erro de predição em cada subgrupo. Cada nó interno da árvore representa uma condição sobre uma variável, enquanto cada folha representa uma predição final (valor médio dos dados que caem naquele nó terminal).

Ao contrário da regressão linear, esse modelo **não assume linearidade** nem independência entre as variáveis preditoras, o que permite capturar **relações não lineares** e **interações complexas** entre os atributos — como, por exemplo, efeitos combinados de temperatura e umidade em diferentes estados.

A árvore de decisão foi selecionada por sua **alta interpretabilidade visual** e por sua **flexibilidade** na modelagem de fenômenos que não seguem padrões estritamente lineares, como é comum em processos ambientais e climáticos. Além disso, tem como aplicação permitir a extração da **importância relativa das variáveis** na formação das decisões, oferecendo *insights* adicionais sobre os fatores mais determinantes para as emissões.

Ambos os modelos foram implementados com a biblioteca scikit-learn, sendo o modelo de árvore complementado por validação cruzada do tipo K-Fold (com k=5k = 5k=5) para mitigar riscos de overfitting e garantir maior robustez na avaliação.

A adoção simultânea de dois algoritmos com naturezas diferentes — um paramétrico e linear, outro não paramétrico e flexível — permitiu contrastar abordagens e avaliar a capacidade de generalização sob diferentes hipóteses, contribuindo para uma compreensão mais abrangente dos dados e da tarefa de predição proposta.

## Percentual da Base de Dados para Treinamento

Para a avaliação dos modelos, a divisão dos dados foi: 70% dos dados para treino e os 30% restantes para teste. Essa partição foi usada tanto na árvore de regressão quanto na regressão linear. A escolha da proporção 70/30 é amplamente adotada em tarefas de regressão, por oferecer um equilíbrio adequado entre a quantidade de dados disponíveis para o aprendizado e a quantidade reservada para validação. Para garantir a reprodutibilidade dos resultados, foi utilizado o mesmo valor de *random state*, 42, em ambas as divisões. Isso assegura que os conjuntos de treino e teste permaneçam consistentes entre diferentes execuções, permitindo uma comparação plausível entre os modelos.

## Pré-processamento

No pré-processamento dos dados, foram utilizadas duas pipelines distintas para os modelos treinados. Ambas incluíram a conversão da variável categórica “\_estado” em variáveis *dummies* por meio da técnica de *one-hot encoding*, utilizando a opção *drop\_first=True* para evitar multicolinearidade perfeita. Essa transformação permitiu representar adequadamente os dados categóricos em formato numérico. Assim, tanto o modelo de árvore de regressão quanto o de regressão linear receberam entradas adequadas para o treinamento.

Além disso, no pipeline de regressão linear, foram removidas algumas variáveis com alta multicolinearidade, a fim de evitar instabilidade nos coeficientes e redundância nas informações fornecidas ao modelo. As variáveis eliminadas foram: temperatura média no ponto de orvalho, umidade relativa média, umidade relativa mínima e velocidade máxima do vento. Para identificar essa multicolinearidade, foram utilizados dois recursos: um mapa de correlação entre as variáveis preditoras, no qual se identificaram os seguintes coeficientes elevados:

* cli\_temp\_orvalho\_med com cli\_umid\_rel\_med: ρ = 0,85
* cli\_umid\_rel\_min\_med com cli\_umid\_rel\_min\_min: ρ = 0,83
* cli\_umid\_rel\_med com cli\_umid\_rel\_min\_med: ρ = 0,84
* cli\_umid\_rel\_med com cli\_temp\_orvalho\_med: ρ = 0,85
* cli\_veloc\_vento\_med com cli\_veloc\_vento\_max: ρ = 0,78

Com base nesses valores, foram removidas as seguintes variáveis explicativas do modelo:

* cli\_temp\_orvalho\_med
* cli\_umid\_rel\_med
* cli\_umid\_rel\_min\_min
* cli\_veloc\_vento\_max

Por fim, apenas na regressão linear, todas as variáveis preditoras foram padronizadas com o *StandardScaler*. Essa etapa é essencial para modelos de regressão linear, pois garante que as variáveis estejam na mesma escala, o que melhora o desempenho numérico do algoritmo e facilita a interpretação dos coeficientes.

## Métricas de Desempenho

Para avaliar o desempenho dos modelos de regressão, foram utilizadas métricas adequadas ao contexto de tarefas de regressão contínua. Entre elas, destaca-se o Coeficiente de Determinação (R²), que mede a proporção da variância da variável dependente explicada pelo modelo. Além disso, foram empregadas métricas de erro, como o Erro Quadrático Médio (MSE), que penaliza mais fortemente grandes desvios; o Erro Absoluto Médio (MAE), que fornece uma medida direta da magnitude média dos erros; e a Raiz do Erro Quadrático Médio (RMSE), que mantém a unidade original da variável alvo e facilita a interpretação prática do erro médio. No caso da árvore de regressão, também foi utilizada a técnica de validação cruzada com cinco dobras (cross-validation), aplicando as mesmas métricas em cada subdivisão para obter uma estimativa mais robusta da performance do modelo. Essas métricas permitem uma análise abrangente da capacidade de generalização e da precisão dos modelos construídos.

TROCAR COM A INTRODUÇÃO

# Obtenção e Tratamento dos Dados

## Justificativa para a Escolha dos Dados/Atributos

Por ser a variável alvo, é indubitável que a **emissão de CO₂** devesse ser um dos atributos. Nesse cenário, como a QP define que as previsões devem ter granularidade mensal e estadual; **ano**, **mês** e **estado** também precisaram ser incluídos, a fim de permitirem a agregação correta dos dados.

Quanto ao escopo de queimadas, também delimitado na QP, optou-se apenas pela **área queimada** e pela **quantidade de focos**, porque quantificam as queimas de forma contínua e discreta, o que fornece uma descrição complementar sobre a aparição desses eventos no espaço geográfico. Foram também considerados outros fatores, a saber: *Fire Radiative Power* (denotado por FRP; mede a intensidade de energia emitida por uma queimada); risco de fogo; e dias sem chuva. O FRP detalha a intensidade de uma queimada, enquanto as demais medidas descrevem fatores que a propiciam. Pela indisponibilidade de dados, todavia, precisaram ser descartados (ao menos nesta primeira *sprint*). Por fim, foram cogitados indicadores socioeconômicos que se relacionam com as queimadas vindas da ação antrópica, mas não foram incluídos por conta da complexidade que trariam.

Quanto ao escopo ambiental, consideraram-se os indicadores disponíveis pelo Instituto Nacional de Meteorologia (INMET). Dessa forma, a inclusão ou exclusão dos atributos dependeu diretamente da disponibilidade e da viabilidade de extração, haja vista que muitas das medições eram horárias e deveriam ser agregadas em valores mensais. Em suma, depois de muitas coletas e reestruturações, retiveram-se **pressão atmosférica**, **umidade relativa do ar**, **velocidade do vento** e **temperatura**, todas as quais se relacionam com a produção de CO₂.

## Fontes dos Dados

Foram utilizadas as seguintes fontes, todas públicas e com dados gratuitos para *download*. Uma vez que todas são organizações governamentais ou civis com compromisso socioambiental e científico, acreditou-se serem confiáveis e de qualidade.

### Dados de emissões

**Banco de Dados do SEEG (Sistema de Estimativas de Emissões de Gases de Efeito Estufa) –** O SEEG, mantido pelo Observatório do Clima [2025], fornece estimativas de emissões de gases de efeito estufa, incluindo CO₂, para diferentes setores e regiões do Brasil, inclusive a Amazônia Legal.

### Dados de queimadas

**MapBiomas – Monitor do Fogo –** O Monitor do Fogo, desenvolvido pelo MapBiomas [2025], oferece uma plataforma interativa de monitoramento de áreas queimadas no Brasil, com base em imagens de satélite processadas automaticamente. O sistema permite a visualização e a análise temporal das ocorrências de fogo em diferentes biomas, com alta resolução espacial e atualizações frequentes.

**Instituto Nacional de Pesquisas Espaciais (INPE) – Programa Queimadas –** O Programa Queimadas do INPE [2025] fornece dados detalhados sobre focos de calor e áreas queimadas no Brasil e na América Latina, com atualizações quase em tempo real. A plataforma **TerraBrasilis** permite acesso a séries históricas desde 1998, além de mapas, estatísticas e análises espaciais por bioma, estado e município, sendo uma referência nacional no monitoramento de incêndios florestais.

### Dados ambientais

**Instituto Nacional de Meteorologia (INMET)** – O INMET [2025] disponibiliza dados meteorológicos oficiais do Brasil, incluindo medições de temperatura, umidade, pressão atmosférica, velocidade do vento e precipitação. Essas informações, obtidas por meio de uma rede nacional de estações, são fundamentais para análises climáticas e ambientais com granularidade temporal variada, incluindo séries históricas mensais.

Para uma visão mais específica, segue o dicionário de fontes:

| Categoria | Espaço | Tempo | Descrição | Fonte |
| --- | --- | --- | --- | --- |
| Clima | município | hora | Fatores climáticos (temperatura, precipitação, umidade relativa do ar, dentre outros) a partir de medições horárias de estações meteorológicas em alguns municípios a partir de 2000. | INMET |
| CO2 | município | ano | Emissões de CO2 em toneladas por estado do Brasil por ano a partir de 2000. | Observatório do Clima |
| Queimadas | estado | mês | Mapeamento mensal de cicatrizes de fogo por estado a partir de 2019. | MapBiomas |
| Queimadas | estado | mês | Série histórica, mapeada de 1985 a 2023, que apresenta as cicatrizes do fogo mensais e por estado. | MapBiomas |
| Queimadas | estado | dia, mês, ano | Fire Radiative Power (FRP), quantidade de focos de queimadas, risco de fogo e dias sem chuva, todos por estado por mês a partir de 2019. | TerraBrasilis |
| Queimadas | estado | mês | Quantidade de focos de queimadas por estado por mês a partir de 1998. | TerraBrasilis |

Tabela 1 – Dicionário de fontes do projeto

## Procedimentos de Limpeza/Transformação/Redução

### Dados de emissões

Foi necessário reformatar o arquivo CSV vindo do SEEG, pois seu formato mostrou-se incompreensível pelo **pandas**. Diante disso, fez-se uso gratuito do chat GPT 4o-mini com o seguinte *prompt*: *Poderia formatar adequadamente o suposto csv que lhe enviarei? Quando fui abri-lo no pandas, os dados não foram processados adequadamente.*

*Outliers*

Quanto a *outliers*, optou-se por deliberadamente não os remover, uma vez que a emissão de dióxido de carbono na atmosfera é sensível à atividade antrópica, de natureza volátil e imprevisível, porque é atrelada a dezenas de variáveis: disposições legais do País, demanda por madeira, governo vigente etc.

*Dados faltantes*

Como a granularidade temporal dos dados a alimentarem o modelo deveria ser mensal, fez-se, sob sugestão da professora Paula Maçaira, a seguinte extrapolação: não tendo sido possível coletar informações mês a mês sobre as emissões, mas apenas ano a ano; decidiu-se replicar o valor anual ao longo dos meses, a fim de que refletissem o comportamento geral — algo semelhante à imputação da média, mas multiplicada pelo número de meses. Dessa forma, espera-se, o modelo ainda será capaz de identificar padrões, se bem que com menor teor de detalhe. Seja como for, ainda se procuram as informações com a devida granularidade, que poderão ser incorporadas ao projeto em futuras *sprints*.

### Dados de queimadas

*Outliers*

Novamente, não foram removidos *outliers*, porque poderiam indicar eventos atípicos, que contribuiriam para o aumento das emissões.

*Dados faltantes*

Quanto à quantidade de focos, não houve dados faltantes. Quanto à área queimada, foram utilizadas duas bases de dados contendo informações por estado e mês: uma principal, abrangendo o período de **1985 a 2023**; e outra complementar, com dados disponíveis entre **2019 e 2025**. Ambas forneceram registros mensais por unidade federativa, permitindo análises temporais e geográficas das queimadas ao longo de quatro décadas.

Durante o processo de integração, foi identificado que a base principal apresentava valores ausentes nos anos de 2019 a 2023. Para lidar com essas lacunas, optou-se por substituí-las por dados correspondentes da base complementar, considerando a mesma data e estado. Essa decisão foi respaldada por uma análise de correlação entre os dois conjuntos, cujo coeficiente foi de aproximadamente **0,52**. Embora não seja elevado, esse valor foi considerado suficiente para justificar o uso de dados observados, ao invés de estimativas com base em médias.

Nos anos anteriores a 2019 (1985 a 2018), em que não havia sobreposição com a base complementar, os valores ausentes foram preenchidos utilizando a **média mensal por estado**, de modo a preservar padrões sazonais regionais.

Além disso, foram incorporados os registros relativos a 2024 e 2025 provenientes da base complementar, dada a **alta correlação** desses valores com as médias históricas da base principal (**0,88**), conferindo robustez à extrapolação temporal do conjunto final.

### Dados ambientais

*Outliers*

Trataram-se os *outliers* após comparar os resultados com e sem o tratamento. Mantendo-os, observaram-se valores incoerentes com a realidade, como temperaturas médias em torno de 12°C, além de um aumento significativo na quantidade de dados faltantes.

*Dados faltantes*

Em relação aos valores ausentes, optou-se por remover a coluna de radiação, que apresentava mais de 70% dos dados faltantes. Consideramos inviável aplicar técnicas de imputação, tanto pela grande quantidade de valores ausentes quanto pela baixa relevância da radiação para as queimadas na Amazônia. A região apresenta alta umidade ao longo do ano, o que dificulta a ocorrência de queimadas por causas naturais, como o acúmulo de radiação solar.

## Tamanho da Base Final

Após o tratamento, resultaram 1025 tuplas com 15 atributos cada, conforme ilustram as figuras abaixo:

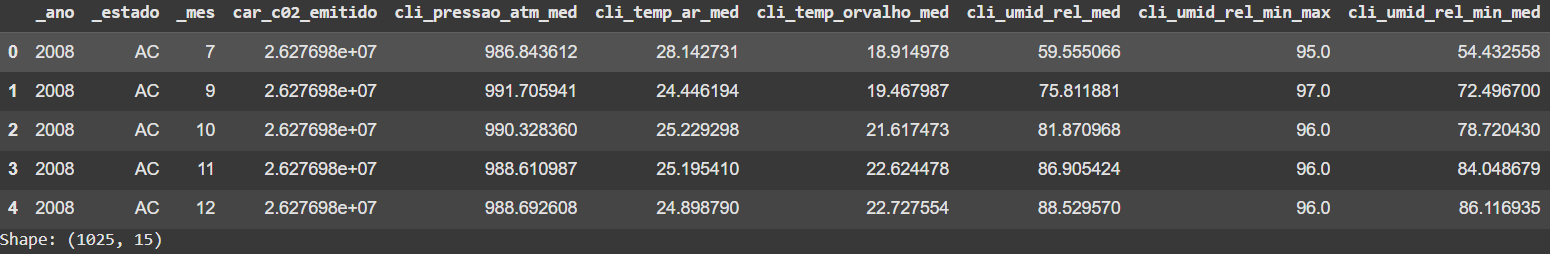


Figura 1 – Primeira metade do *dataset* após o tratamento

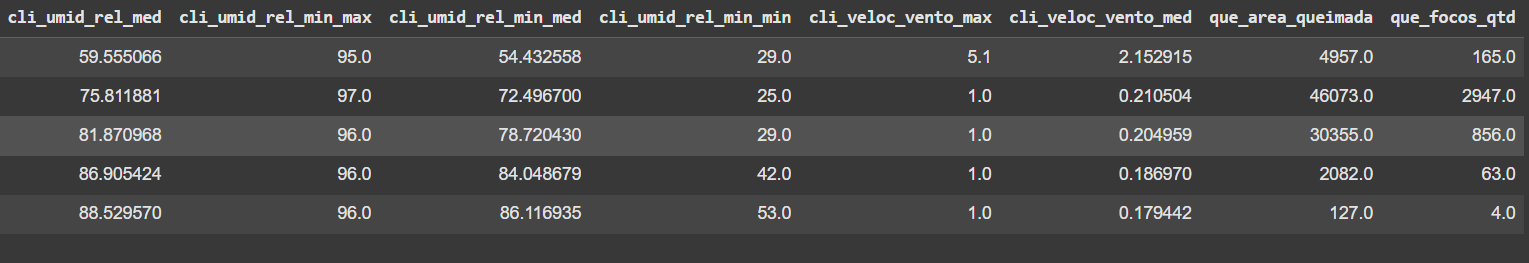
**

Figura 2 – Segunda metade do *dataset* após o tratamento

## Dicionário de Dados

A seguir, produziu-se um resumo do dicionário de dados do projeto, cuja versão completa encontra-se em [dicionarioDados](https://docs.google.com/spreadsheets/d/1hZrFiSA9nULC0x-QHo0SyI0-1PRj0odciN_ByROKZt8/edit?usp=sharing). Os atributos que compõem a chave primária do *dataset* foram grifados em vermelho.

| Nome | Tipo | Descrição | Unidade de medida | Fonte |
| --- | --- | --- | --- | --- |
| \_ano | Int32 | Ano segundo o calendário cristão. | ano | Não se aplica |
| \_estado | string | Nome do estado da Amazônia Legal. | adimensional | Não se aplica |
| \_mes | Int32 | Mês do ano segundo o calendário cristão. Janeiro relaciona-se com 1, fevereiro com 2 e assim em diante. | mês | Não se aplica |
| car\_c02\_emitido | Float64 | Quantidade de dióxido de carbono emitido na atmosfera. Se for negativa, considera-se que tenha havido absorção. | toneladas (t) | SEEG |
| cli\_pressao\_atm\_med | Float64 | Média das medições mensais de pressão atmosférica de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET). | mb (milibar) | INMET |
| cli\_temp\_ar\_med | Float64 | Média das medições mensais de temperatura do ar de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET). | °C | INMET |
| cli\_temp\_orvalho\_med | Float64 | Média das medições mensais de temperatura do ponto de orvalho de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET). | °C | INMET |
| cli\_umid\_rel\_med | Float64 | Média das medições mensais de umidade relativa do ar de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET). | percentual (%) | INMET |
| cli\_umid\_rel\_min\_max | Float64 | Máxima dentre as mínimas mensais de umidade relativa do ar de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET). | percentual (%) | INMET |
| cli\_umid\_rel\_min\_med | Float64 | Média das mínimas mensais de umidade relativa do ar de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET). | percentual (%) | INMET |
| cli\_umid\_rel\_min\_min | Float64 | Mínima dentre as mínimas mensais de umidade relativa do ar de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET). | percentual (%) | INMET |
| cli\_veloc\_vento\_max | Float64 | Máxima dentre as medidas mensais de velocidade do ar de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET). | m/s | INMET |
| cli\_veloc\_vento\_med | Float64 | Média das medidas mensais de velocidade do ar de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET). | m/s | INMET |
| que\_area\_queimada | Float64 | Número de hectares de área queimada. | hectare (Ha) | MapBiomas |
| que\_focos\_qtd | Float64 | Quantidade de focos de queimada. Se bem que, em essência, seja um valor inteiro, foi considerado um domínio de ponto flutuante por conta de imputação de dados. | adimensional | TerraBrasilis |

Tabela 2 – Dicionário de dados do projeto

# Análise Exploratória dos Dados

Para sistematizar a análise, adotou-se, a nível de código, o seguinte critério: *Se uma dada variável apresentar tipo numérico (***int***,***float***, dentre outros), automaticamente será* ***quantitativa****. A cardinalidade será, então, dada pela do tipo: inteiros são contáveis (****discretos****), e pontos flutuantes representam intervalos reais, não contáveis (****contínuos****). Caso contrário, assume-se ser qualitativa, e a subcategoria é avaliada*ad hoc*por análise humana.*

Diante disso, vale ressaltar que, embora **\_mes** fosse originalmente uma variável qualitativa ordinal, com domínio **(janeiro, fevereiro, ...)**, com a codificação real para **(1, 2, ...)**, pode-se interpretá-la como a quantidade de meses desde o início do ano, algo quantitativo discreto. Quanto a **\_estado**, como não há ordem explícita, segue que é quantitativa nominal.

Assim, as variáveis terão a seguinte classificação:

1. **\_ano:** quantitativo discreto;
2. **\_estado**: qualitativo nominal;
3. **\_mes**: quantitativo discreto;
4. **car\_c02\_emitido**: quantitativo contínuo;
5. **cli\_pressao\_atm\_med**: quantitativo contínuo;
6. **cli\_temp\_ar\_med**: quantitativo contínuo;
7. **cli\_temp\_orvalho\_med**: quantitativo contínuo;
8. **cli\_umid\_rel\_med**: quantitativo contínuo;
9. **cli\_umid\_rel\_min\_max**: quantitativo contínuo;
10. **cli\_umid\_rel\_min\_med**: quantitativo contínuo;
11. **cli\_umid\_rel\_min\_min**: quantitativo contínuo;
12. **cli\_veloc\_vento\_max**: quantitativo contínuo;
13. **cli\_veloc\_vento\_med**: quantitativo contínuo;
14. **que\_area\_queimada**: quantitativo contínuo;
15. **que\_focos\_qtd**: quantitativo contínuo.

## Estatísticas Descritivas

A análise descritiva visa entender o comportamento das variáveis quantitativas mais relevantes para o problema, como emissão de CO₂, focos e área de queimadas, além de variáveis ambientais. Foram utilizadas medidas de tendência central, dispersão, posição e forma da distribuição para detectar padrões, outliers e características que podem impactar a modelagem.

1. Tendência Central

Foram calculadas as seguintes métricas para representar o comportamento central dos dados:

* Média: Valor médio da distribuição.
* Mediana: Valor que divide a amostra em duas partes iguais (50% dos dados).
* Moda: Valor mais frequente (ou múltiplas modas, se aplicável).

Análises:

A “car\_c02\_emitido” apresenta média (1.075e+08) significativamente maior que a mediana (5.667e+07), sugerindo assimetria à direita (valores extremos elevados).

Para “que\_focos\_qtd”, a mediana (26.0) é muito inferior à média (571.35), indicando alta concentração de valores baixos com outliers positivos.

2. Dispersão

As métricas de variabilidade incluem:

* Amplitude: Diferença entre máximo e mínimo.
* Variância e Desvio Padrão: Medem a dispersão em torno da média.
* Coeficiente de Variação (CV): Razão entre desvio padrão e média (útil para comparar variáveis em escalas distintas).

Análises:

“que\_area\_queimada” e “que\_focos\_qtd” apresentam CV > 2, indicando alta dispersão relativa (dados heterogêneos).

“\_ano” tem CV ≈ 0.0026, confirmando baixa variabilidade (dados concentrados em um período específico).

3. Posição (Quartis)

Os quartis (Q1, Q2=mediana, Q3) dividem os dados em intervalos de 25%:

Exemplo: Para “cli\_temp\_ar\_med”, 25% dos valores estão abaixo de 25.46°C (Q1), enquanto 75% estão abaixo de 27.30°C (Q3).

Análises:

Em cli\_veloc\_vento\_med, a diferença entre Q3 (1.479) e Q1 (0.824) sugere dispersão moderada na velocidade média do vento.

4. Forma da Distribuição

Assimetria (Coeficiente de Fisher-Pearson)

Valores próximos de zero indicam simetria, enquanto:

Assimetria positiva (cauda à direita):

“car\_c02\_emitido” (2.106), “que\_focos\_qtd” (2.802), e “que\_area\_queimada” (5.136) mostram outliers elevados.

Assimetria negativa (cauda à esquerda):

“cli\_pressao\_atm\_med” (-2.161) e “cli\_temp\_orvalho\_med” (-2.164) têm concentração de valores altos.

Curtose (Coeficiente de Fisher)

Mede o "achatamento" da distribuição:

Leptocúrtica (curtose > 0): Caudas pesadas e pico agudo.

“cli\_umid\_rel\_min\_max” (45.515) e “que\_area\_queimada” (39.861) indicam outliers extremos.

Platicúrtica (curtose < 0): Distribuição mais achatada (ex.: “\_ano”, “\_mes”).

Observação: Essas métricas foram calculadas usando Series.skew() e Series.kurtosis() do pandas.

Conclusão

As estatísticas descritivas revelam padrões importantes, como:

Variáveis ambientais (ex.: cli\_temp\_ar\_med) tendem a ser simétricas (assimetria próxima de zero).

Dados de emissão de CO₂ e queimadas apresentam assimetria positiva e curtose elevada, confirmando a presença de outliers. Isso sugere a necessidade de transformações (ex.: logarítmica) ou métodos robustos para modelagem.

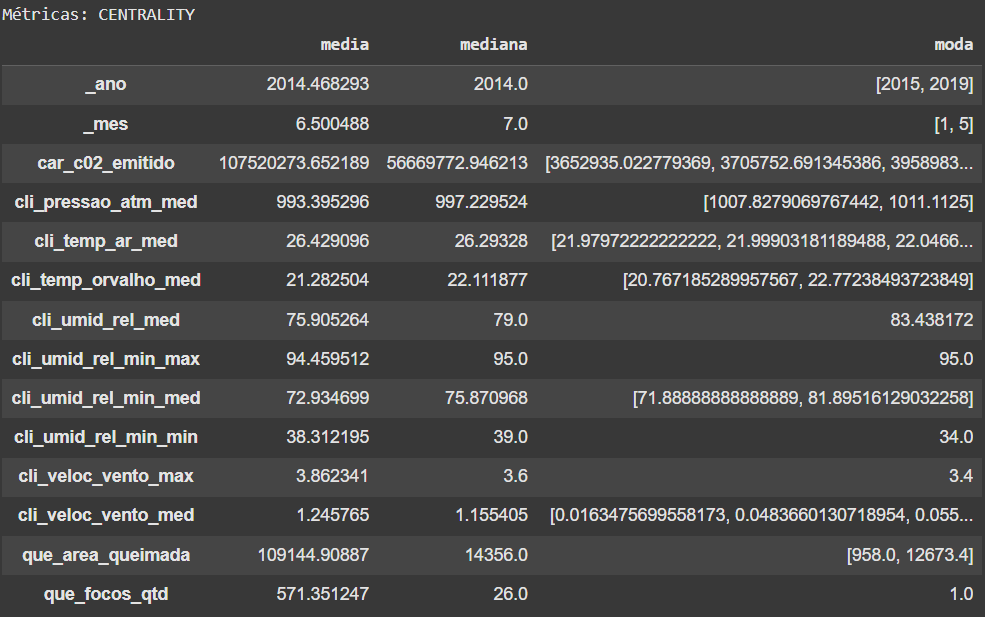


Figura 3 – Medidas de centralidade das variáveis quantitativas

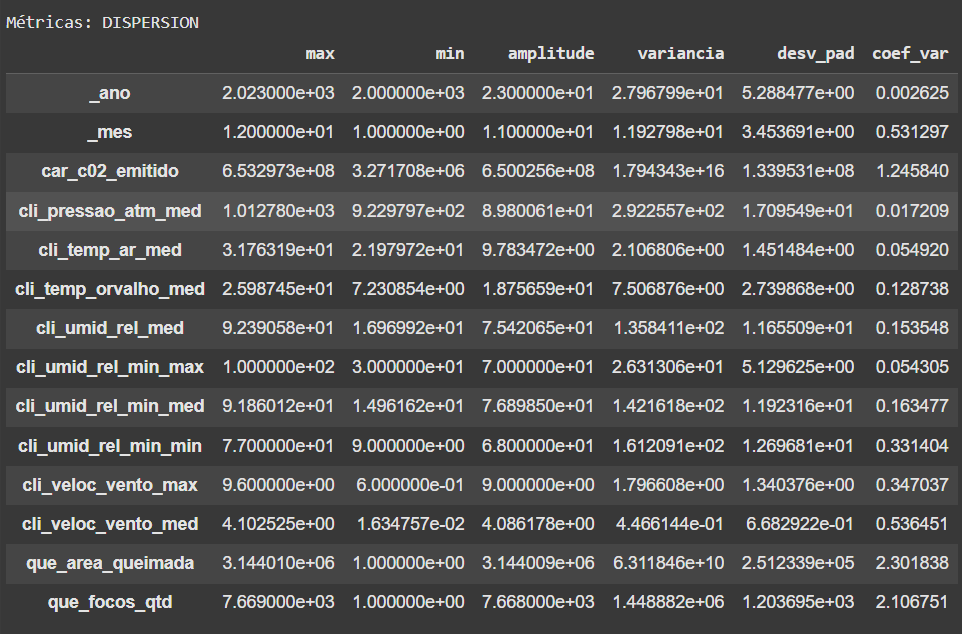


Figura 4 – Medidas de dispersão das variáveis quantitativas

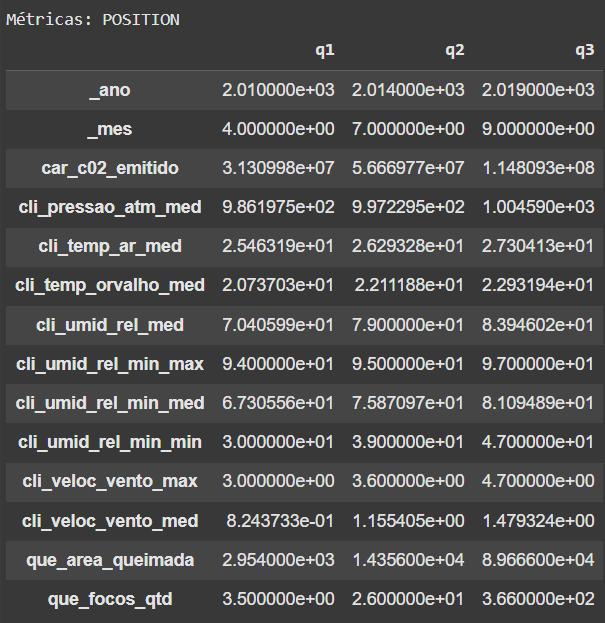


Figura 5 – Medidas de posição das variáveis quantitativas



Figura 6 – Medida de assimetria das variáveis quantitativas

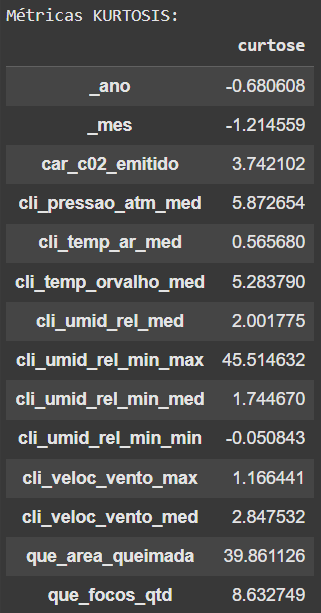


Figura 7 – Medida de curtose das variáveis quantitativas

## Visualizações de Distribuição

O gráfico KDE revela a forma da distribuição das emissões de CO₂: observa-se que a maior densidade de probabilidade ocorre para valores inferiores a 200 milhões de toneladas por mês, o que está alinhado com a média (107,5 milhões) e a mediana (56,7 milhões) calculadas anteriormente. A distribuição apresenta assimetria positiva, evidenciada pela cauda alongada à direita, que confirma a presença de valores extremamente altos (acima de 600 milhões), embora com baixa probabilidade. Além disso, o pico estreito e as caudas relativamente pesadas são compatíveis com o coeficiente de curtose positivo (3,74), caracterizando uma distribuição leptocúrtica, ou seja, com maior concentração de dados próximos à média em comparação com uma distribuição normal. Observa-se também que a densidade é próxima de zero para emissões abaixo de zero, o que descarta a possibilidade de absorção líquida de CO₂ nos dados analisados. Por fim, vale destacar que o KDE suaviza os dados, o que pode mascarar variações locais importantes que seriam visíveis em um histograma.

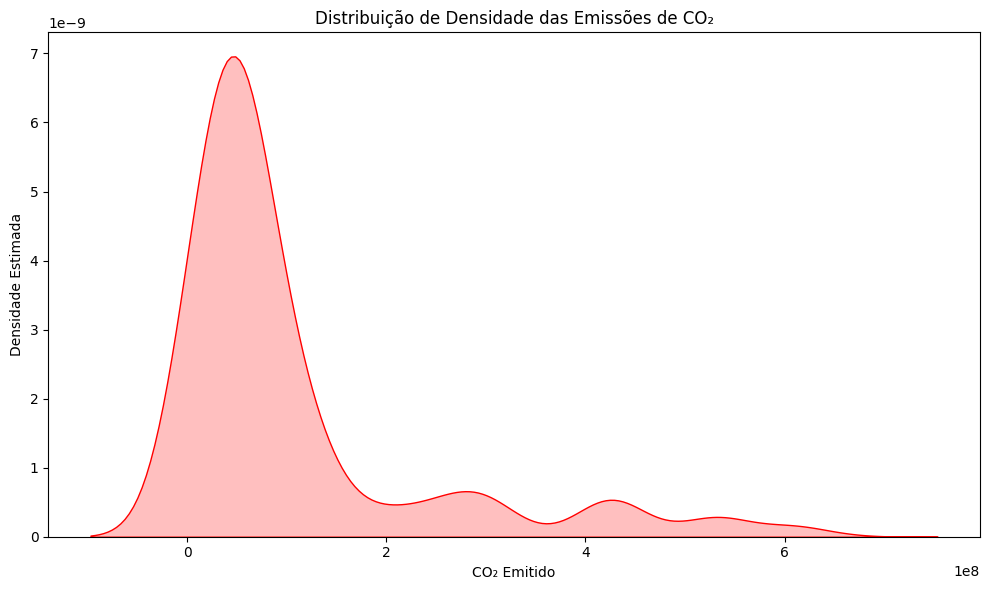


Figura 8 – Gráfico de densidade (KDE) da emissão de CO₂

O histograma complementa o gráfico de KDE ao mostrar a distribuição real das emissões de CO₂ em intervalos discretos. A maior parte das observações, correspondendo a mais de 80% dos dados segundo a contagem de *bins*, está concentrada abaixo de 200 milhões de toneladas, enquanto pouquíssimos valores ultrapassam os 600 milhões, apenas um ou dois *bins* apresentam frequências próximas de zero nessa faixa. Visualmente, observa-se uma assimetria positiva, com a barra mais alta localizada à esquerda e uma cauda longa e esparsa à direita, o que reforça o coeficiente de assimetria de 2,106. Não há registros em *bins* abaixo de zero, o que confirma a ausência de valores negativos e valida a conclusão já apontada pelo KDE. A forma do histograma também se alinha às métricas estatísticas discutidas na Seção 4.1, especialmente no que diz respeito à assimetria e à curtose. Além disso, a amplitude de 650 milhões e o desvio padrão elevado (1,34 × 10⁸) ajudam a explicar a dispersão observada nas extremidades da distribuição.

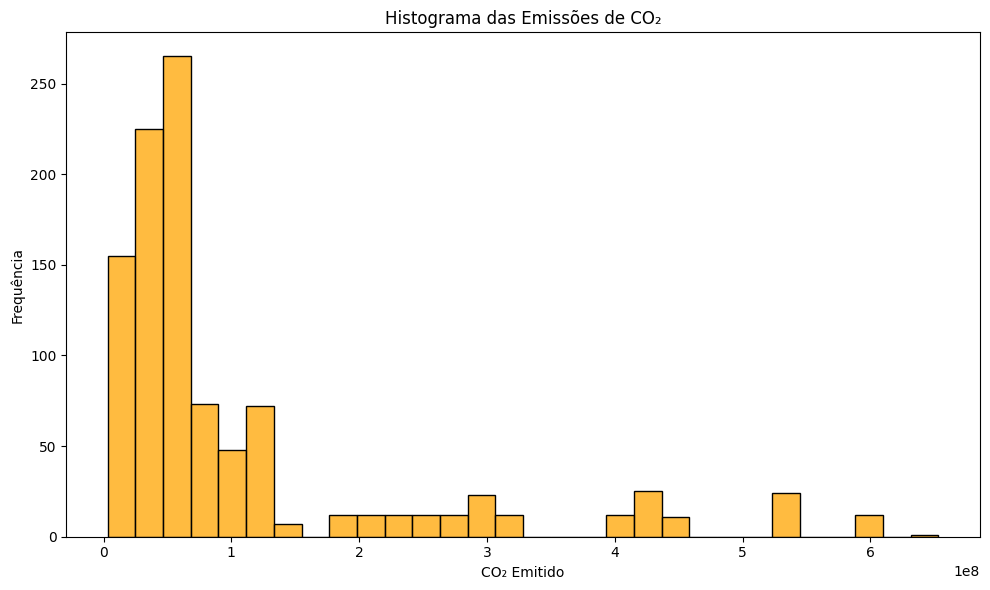


Figura 9 – Histograma da emissão de CO₂

## Análise de Correlação

Consideram-se como **pertinentes as correlações  com valores absolutos maiores ou iguais a 0.70**, segundo o critério usual para indicar **correlação forte**. Correlações **fracas ou próximas de zero** (entre -0.05 e 0.05) também são destacadas por sua **irrelevância estatística**, assim como aquelas **próximas do corte de 0.70**, que podem indicar relações latentes interessantes.

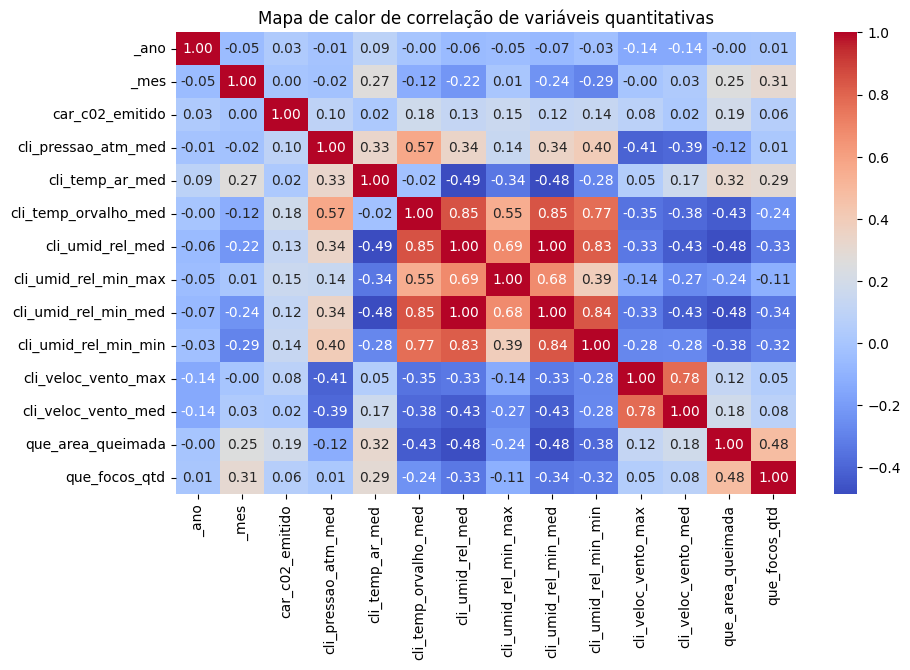


Figura 10 – Mapa de calor de correlação de variáveis quantitativas

**Análise de Correlação**

**1. Correlações Fracas (|ρ| ≤ 0.05)**

**As seguintes variáveis apresentam correlação insignificante, indicando ausência de relação linear:**

**\_ano ↔ cli\_pressao\_atm\_med (ρ = -0.01): A pressão atmosférica não variou sistematicamente ao longo dos anos.**

**\_mes ↔ car\_c02\_emitido (ρ = 0.00): Não há evidência de sazonalidade nas emissões de CO₂.**

**\_ano ↔ cli\_temp\_orvalho\_med (ρ = -0.00): A temperatura de orvalho manteve-se estável temporalmente.**

**cli\_veloc\_vento\_med ↔ car\_c02\_emitido (ρ = 0.02): A velocidade média do vento não influencia as emissões.**

**Observação: Correlações próximas de zero não descartam relações não-lineares ou contextos específicos não capturados.**

**2. Correlações Moderadas (0.05 < |ρ| < 0.70)**

**Relações parciais que exigem contextualização:**

**que\_area\_queimada ↔ car\_c02\_emitido (ρ = 0.19):**

**A baixa correlação contraria a expectativa teórica, possivelmente devido à dispersão temporal (dados anuais imputados para meses). Quando agregados por estado, a correlação sobe para 0.57, revelando distorções na escala mensal.**

**que\_area\_queimada ↔ que\_focos\_qtd (ρ = 0.48):**

**Indica que mais focos tendem a aumentar a área queimada, mas a variabilidade sugere outros fatores em jogo (ex.: tipo de vegetação, ações de combate).**

**cli\_umid\_rel\_med ↔ cli\_umid\_rel\_min\_max (ρ = 0.69):**

**Reflete a coerência interna dos dados climáticos: meses úmidos têm mínimas diárias menos extremas.**

**3. Correlações Fortes (|ρ| ≥ 0.70)**

**Padrões robustos e esperados:**

**Variáveis climáticas:**

**cli\_veloc\_vento\_max ↔ cli\_veloc\_vento\_med (ρ = 0.78): Velocidades máximas e médias do vento são diretamente proporcionais.**

**cli\_umid\_rel\_med ↔ cli\_temp\_orvalho\_med (ρ = 0.85): A temperatura de orvalho é função da umidade, como esperado na termodinâmica.**

**Umidade relativa:**

**Correlações acima de 0.77 entre médias e mínimas (ex.: cli\_umid\_rel\_med ↔ cli\_umid\_rel\_min\_min) confirmam a estabilidade espacial da umidade.**

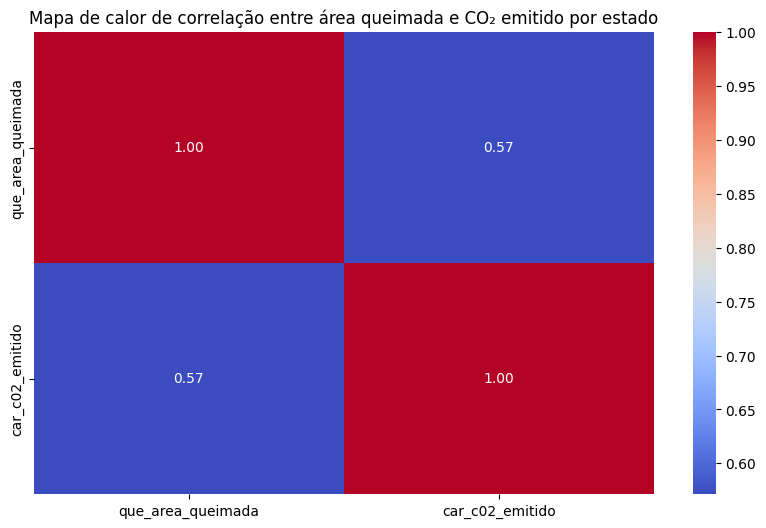


Figura 11 – Correlação entre área queimada e CO₂ emitido (por estado)

## Gráficos de Dispersão

Este gráfico de dispersão mostra a relação entre a quantidade de focos de queimada e a emissão de CO₂ nos diferentes estados da Amazônia Legal. De forma geral, não se observa uma correlação linear forte e direta entre essas duas variáveis. Embora haja uma concentração de pontos com baixos valores tanto para focos quanto para emissão de CO₂, percebe-se uma grande dispersão nos dados.

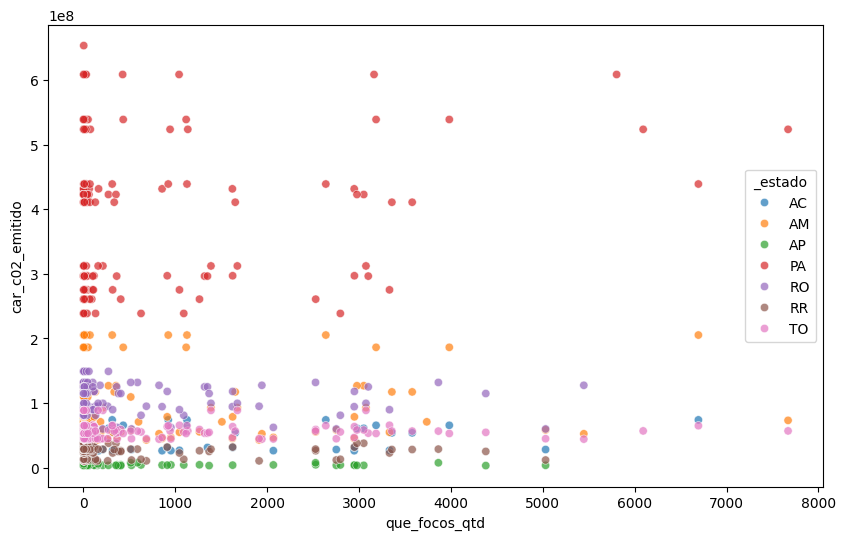


Figura 12 – Focos de queimada *vs* CO₂ emitido

O gráfico de dispersão revela uma relação não linear entre o número de focos de queimadas e as emissões de CO₂ nos estados da Amazônia Legal, evidenciando padrões distintos entre os estados. O Pará, por exemplo, apresenta pontos com poucos focos, mas emissões elevadas, o que pode indicar queimadas de alta intensidade, presença de grande biomassa queimada por foco ou até outras fontes de emissão não diretamente associadas aos focos detectados por satélites. Já o Amazonas e o Acre mostram emissões mais baixas, mesmo com alguma variação no número de focos, possivelmente devido a menor biomassa disponível, queimadas menos intensas ou maior umidade da vegetação, que reduz a eficiência da queima. Esses padrões sugerem que a quantidade de focos, por si só, não explica adequadamente as emissões observadas. Fatores contextuais, como o tipo de vegetação, umidade do solo, estágio do desmatamento e até limitações na detecção por satélite — como a cobertura de nuvens — podem exercer influência significativa sobre os dados. Assim, a relação entre focos de queimadas e emissões de CO₂ se mostra complexa e mediada por múltiplas variáveis ambientais e metodológicas, o que reforça a necessidade de análises complementares com dados sobre biomassa, cobertura vegetal e condições climáticas locais.

Este outro gráfico revela a relação entre a quantidade de focos de queimada e a área queimada nos estados da Amazônia Legal. De maneira geral, observa-se uma correlação positiva, indicando que um maior número de focos tende a estar associado a uma maior área queimada, embora essa relação apresente uma considerável variabilidade.

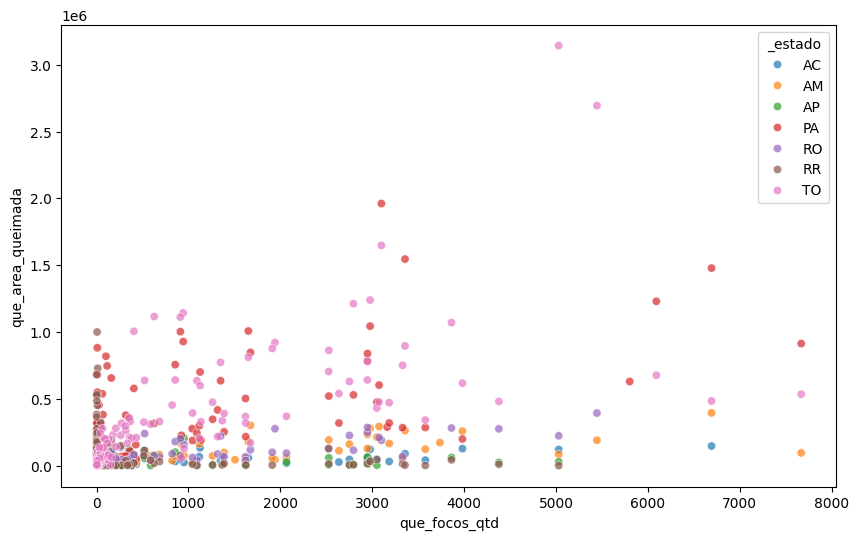


Figura 13 – Área queimada *vs* focos de queimada

O segundo gráfico revela uma correlação positiva entre o número de focos de queimadas e a área queimada, mas essa relação apresenta variações significativas entre os estados da Amazônia Legal. Observa-se uma maior concentração de pontos na parte inferior esquerda do gráfico, indicando que a maioria dos eventos envolve poucos focos e resulta em áreas queimadas relativamente pequenas. No entanto, ao se analisar por estado, padrões distintos emergem. O Pará e o Tocantins, por exemplo, apresentam ocorrências extremas, com muitos focos e áreas queimadas extensas, o que pode refletir dinâmicas associadas à fronteira agrícola, à presença de vegetação mais inflamável ou à incidência de secas severas. Por outro lado, estados como o Acre e o Amapá concentram-se em eventos de menor magnitude, possivelmente em função de políticas ambientais mais restritivas, menor pressão de desmatamento ou características geográficas específicas. A dispersão dos dados, portanto, evidencia que o número de focos, embora relacionado à área queimada, não é um indicador suficiente por si só. Fatores como o tipo de cobertura vegetal, a duração dos incêndios e condições climáticas locais também influenciam significativamente os resultados. Isso sugere a necessidade de análises mais detalhadas, que incorporem variáveis de controle — como bioma, precipitação ou uso do solo — e investiguem casos atípicos, como áreas com grandes queimadas e poucos focos detectados, que podem indicar subnotificação ou falhas nos métodos de medição.

## Análise de Dados Categóricos

Como visto ao início da seção, a única variável categórica é **\_estado**, cuja análise foi feita por meio da seguinte tabela de frequência:

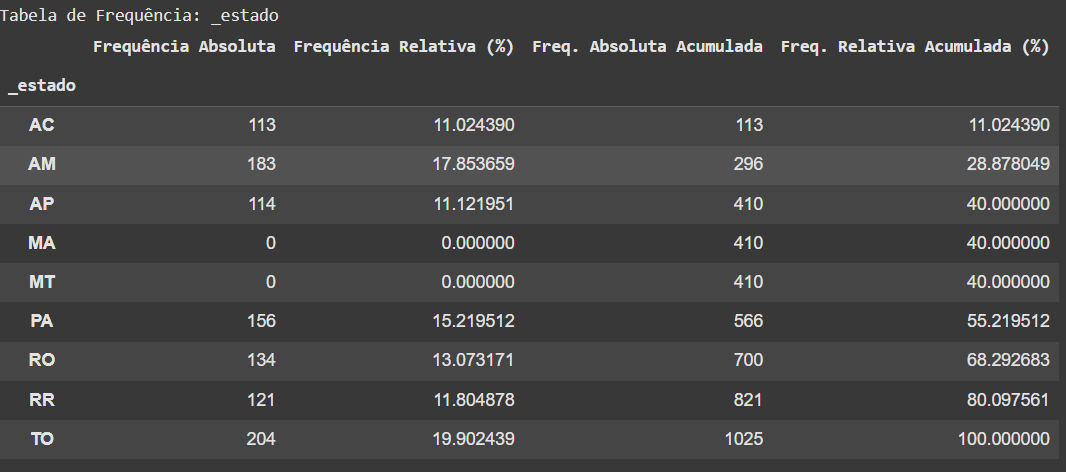
**

Figura 14 – Tabela de frequência de **\_estado**

Verifica-se que **Amazonas** e **Tocantins** são os estados mais representados neste *dataset*, enquanto **Maranhão** e **Mato Grosso** sequer foram representados, cuja ausência explica-se pela dificuldade em coletar suas informações climáticas, o que impossibilitou sua aparição no *dataset* final. Seja como for, uma vez que a quantidade de dióxido de carbono emitida na atmosfera não deveria depender diretamente da posição geográfica, mas sim da área efetivamente queimada, provavelmente não serão necessárias nem a sub nem a sobreamostragem para compensar o desbalanço da amostra. Apesar disso, nas próximas *sprints*, pretende-se coletar mais dados, com o intuito de suavizar essas assimetrias.

## Análise Temporal

Este primeiro gráfico apresenta a evolução anual da emissão de CO₂ na Amazônia Legal entre os anos de 2000 e 2023.

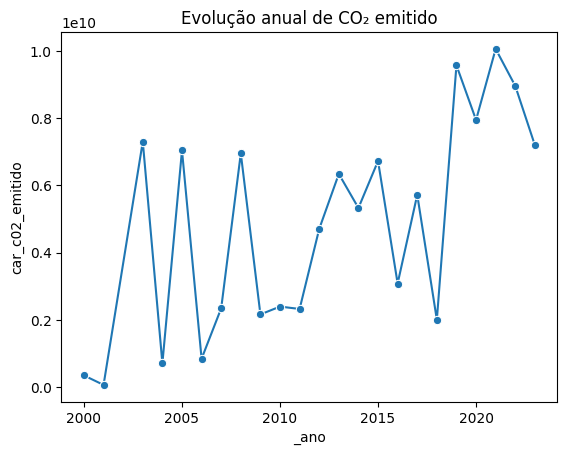


Figura 15 – Evolução anual do CO₂ emitido

Há um padrão bastante oscilante ao longo do período, mas com uma tendência geral de crescimento, principalmente a partir de 2011. Nos primeiros anos da série, entre 2000 e 2003, os valores de emissão eram bastante oscilantes. Em 2003, há um pico, ultrapassando 7 bilhões de toneladas de CO₂ emitidas, seguido por uma queda acentuada em 2004, continuando com altas oscilações até 2010. A partir de 2011, nota-se um crescimento mais contínuo e consistente nas emissões, com menores quedas abruptas e mais anos consecutivos de aumento. Entre 2013 e 2017, os valores se mantêm em patamares intermediários, entre 5 e 7 bilhões de toneladas, com leves oscilações. A partir de 2018, no entanto, há um salto significativo, com os anos de 2019 e 2021 registrando os maiores valores da série, próximos ou acima de 10 bilhões de toneladas de CO₂ emitidos. Após esse pico, observa-se uma leve queda, mas os valores permanecem bastante altos até 2023.

Este outro gráfico mostra como a área queimada na Amazônia Legal evoluiu entre os anos de 2000 e 2023.

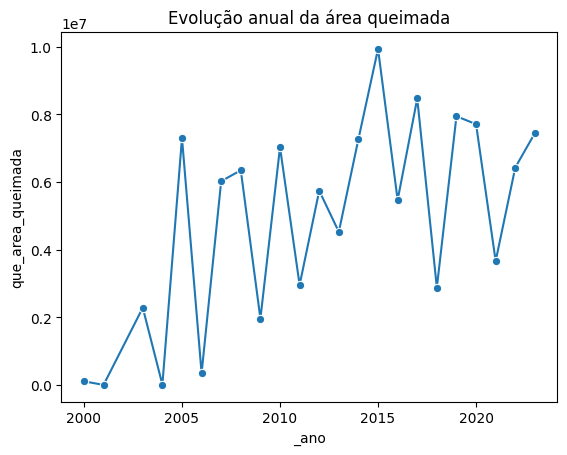


Figura 16 – Evolução anual da área queimada

É possível perceber uma tendência de aumento ao longo do tempo, com várias oscilações de um ano para o outro, que podem ter causas climáticas, políticas ambientais e econômicas. No começo, entre 2000 e 2003, os números eram mais baixos, mas a partir de 2004 já começa um crescimento forte, mesmo com altos e baixos. O ponto mais crítico foi em 2015, quando mais de 10 milhões de hectares foram queimados. Depois disso, os valores continuaram altos, variando bastante a cada ano. Entre 2020 e 2023, houve uma queda seguida de novo aumento. Mesmo nos anos a partir de 2010 que houve queda, os valores ainda são muito maiores do que os do começo da série.

Por fim, tem-se a sobreposição das séries mensais de CO₂, área queimada e quantidade de focos.

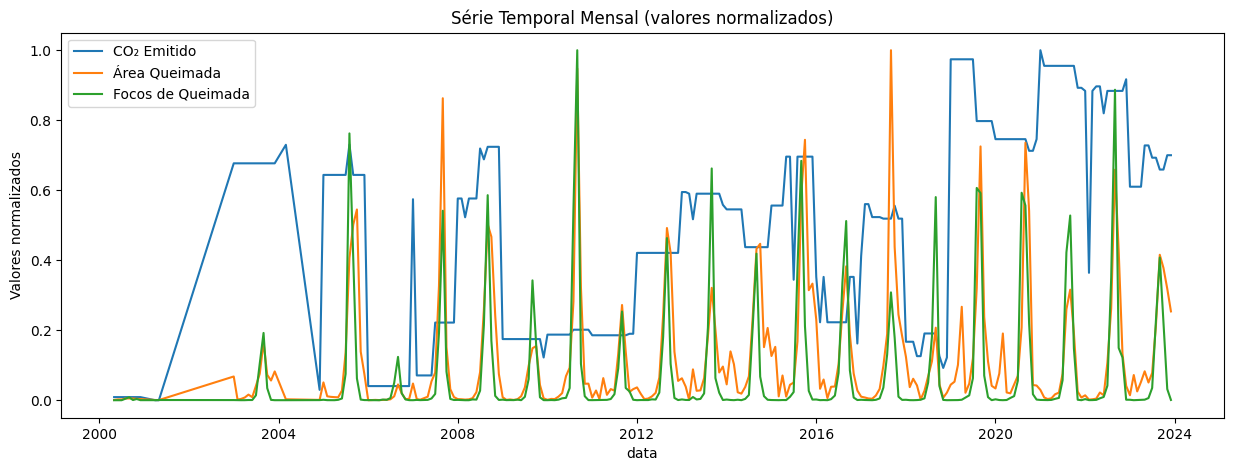


Figura 17 – Série temporal mensal de CO₂ emitido, área queimada e focos

Visualmente, há correlação entre a quantidade de área queimada e a quantidade de focos, haja vista que muitos de seus picos nas séries temporais coincidem. Apesar disso, é nítido que os picos menores de área queimada não costumam ser acompanhados por picos menores da curva de focos, que se mantém plana. Isso pode estar relacionado ao fato de que grandes focos, embora não numerosos, produzam vasta área de queima.

Quanto às emissões de carbono, a imputação dos valores anuais a todos os meses parece ter introduzido demasiada distorção, de modo que a série de CO₂ mantenha-se em alta grande parte do tempo, mesmo que desacompanhada pelas demais. Diante disso, para evitar interpretações enganosas, deve-se procurar dados propriamente mensais sobre o tópico, ainda que disponíveis em um intervalo mais curto que o escopo do projeto.

## Análise Geoespacial

Neste gráfico, quanto maior a intensidade da cor, maior é a área queimada em cada estado.

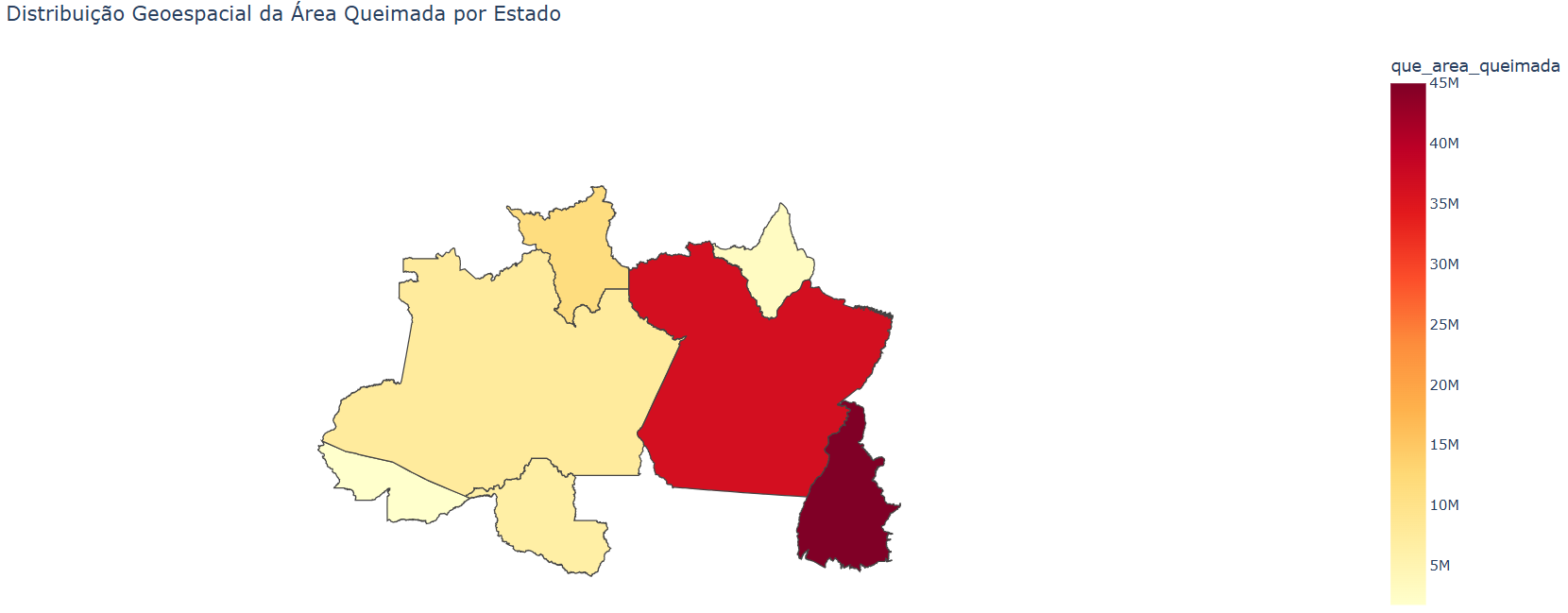


Figura 18 – Área queimada acumulada por estado

Observa-se que o Tocantins, seguido do Pará, apresenta as maiores extensões de áreas queimadas. Em contraste, o Acre e o Amapá registram as menores quantidades. Já os estados do Amazonas, Rondônia e Roraima situam-se em uma posição intermediária. Isso é condizente com a realidade, pois o Pará e o Tocantins fazem parte da chamada Fronteira Agrícola do Brasil, onde há intensa expansão de áreas para agricultura e pastagens. O fogo é usado como uma técnica barata para desmatar e “limpar” o terreno. Além disso, o Tocantins enfrenta anualmente um período de estiagem que se inicia em maio, caracterizado por altas temperaturas e baixa umidade relativa do ar. Essas condições tornam a vegetação mais suscetível ao fogo. Ademais, no Pará há grandes áreas de florestas sendo invadidas ilegalmente para fins de especulação fundiária, com o uso do fogo para marcar territórios e abrir áreas. Por fim, vale ressaltar que ausência de dados sobre Maranhão e Mato Grosso não implica que esses estados não tenham áreas queimadas.

## Insights Importantes

A partir das análises acima, constata-se que, ora por subrepresentação (MA e MT), ora por imputação (emissões mensais de CO₂), distorções e enganos podem ser produzidos, como o Maranhão não ter área queimada. Apesar disso, o problema mais premente parece ser o da emissão de carbono, a qual, além de variável-alvo, mostrou suas distorções diante da discrepância entre a correlação com **que\_area\_queimada** antes e depois da agregação por estado. Ademais, algumas correlações mostraram-se surpreendentemente baixas, a saber:

* **que\_area\_queimada ↔ que\_focos\_qtd = 0.48;**
* **que\_area\_queimada ↔ cli\_temp\_ar\_med = 0.03**.

Tal observação revela o fato de que o senso comum pode ser enganoso em algumas situações: com efeito, um pequeno número de grandes focos pode ser mais impactante que um grande número de pequenos. Por outro lado, expectativas confirmaram-se diante das fortes correlações entre variáveis “cognatas”, como **cli\_umid\_rel\_min\_med** e **cli\_umid\_rel\_min\_min**, o que pode demandar seleção de *features* no ciclo interno de pesquisa com o intuito de evitar a *Curse of Dimensionality*.

Por fim, a análise geoespacial, que permitiu relacionar a quantidade de área queimada à expansão da fronteira agrícola, aponta que não apenas os dados climático-ambientais podem ser úteis, mas também os socioeconômicos. Diante do exposto, conclui-se, em suma, que o universo sugerido pelos dados é um mero reflexo da realidade, que tenderá a ser mais ou menos fiel conforme a qualidade e a limpeza do espelho.

# *Resultados Parciais*

# *5.1 Regressão Linear Múltipla*

A **Regressão Linear Múltipla** foi adotada após a constatação de que a variável “que\_area\_queimada”, isoladamente, não apresentava capacidade preditiva satisfatória em relação às emissões de CO₂ na atmosfera. A aplicação de uma Regressão Linear Simples com essa única variável resultou em um **coeficiente de determinação (R²) negativo**, evidenciando desempenho inferior ao de um modelo trivial que prediz a média da variável-alvo para todos os casos.

Adicionalmente, as métricas de erro, como o **Erro Absoluto Médio (MAE)** e o **Erro Quadrático Médio (MSE)**, apresentaram valores **significativamente superiores à variância da variável dependente**, o que reforça o baixo ajuste do modelo e indica que a variabilidade das emissões não foi adequadamente capturada.

Diante disso, optou-se por uma modelagem múltipla, na qual se buscou estimar a emissão mensal de CO₂ como função conjunta de diversos fatores ambientais e indicadores de queimadas. Foram utilizadas **todas as variáveis remanescentes no dataset**, com exceção daquelas removidas por apresentarem **alta multicolinearidade**, condição que compromete a estabilidade e a interpretabilidade dos coeficientes estimados.

As demais variáveis — incluindo indicadores climáticos como temperatura média, umidade mínima média e velocidade média do vento, além de medidas diretas relacionadas às queimadas — foram mantidas e utilizadas na modelagem.

Conforme mencionado na primeira seção, o particionamento da base de dados foi realizado na proporção **70% para treinamento** e **30% para teste**, permitindo a avaliação do modelo em dados não utilizados no ajuste. O modelo final apresentou desempenho substancialmente superior ao da regressão simples, evidenciando maior aderência aos dados e melhor capacidade explicativa.

*5.1.1 Desempenho dos Algoritmos: Regressão Linear Múltipla*

Com o modelo de Regressão Linear Múltipla, nós obtemos as seguintes métricas:

**MSE (Erro Quadrático Médio):** 1,74 × 10⁵ toneladas²

**R² (Coeficiente de Determinação):** 0,9138

**MAE (Erro Absoluto Médio):** 2,77 × 10⁷ toneladas

Com base no modelo de Regressão Linear Múltipla, observamos um excelente desempenho preditivo, com um R² de 0,9138, indicando que cerca de 91% da variação nas emissões de CO₂ é explicada pelas variáveis do modelo, como área queimada e fatores climáticos. Apesar disso, o erro absoluto médio (MAE) de 27,7 milhões de toneladas ainda é elevado, sugerindo impacto de valores extremos ou necessidade de ajustes no modelo. O MSE relativamente baixo indica possível inconsistência de escala ou necessidade de verificação dos dados. Em resumo, o modelo é eficaz na explicação do fenômeno, mas pode ser aprimorado com tratamentos de outliers e ajustes nas variáveis para reduzir os erros absolutos.

5.1.2.  *Análise das Variáveis*

(Avaliar a importância das variáveis para o algoritmo. Apresente visualizações que ajudem a interpretar a contribuição de cada variável para a performance do modelo.)

A análise da importância das variáveis no modelo de regressão linear múltipla foi realizada com base nos coeficientes estimados e nos respectivos p-valores obtidos pelo pacote *Statsmodels*. Como os dados foram previamente padronizados com o *StandardScaler*, os coeficientes da regressão passaram a representar o impacto relativo de cada variável na predição da emissão de CO₂, permitindo uma comparação direta entre variáveis com diferentes escalas originais. Nesse contexto, as variáveis com maior magnitude absoluta nos coeficientes indicam maior influência no modelo, desde que também apresentem significância estatística.

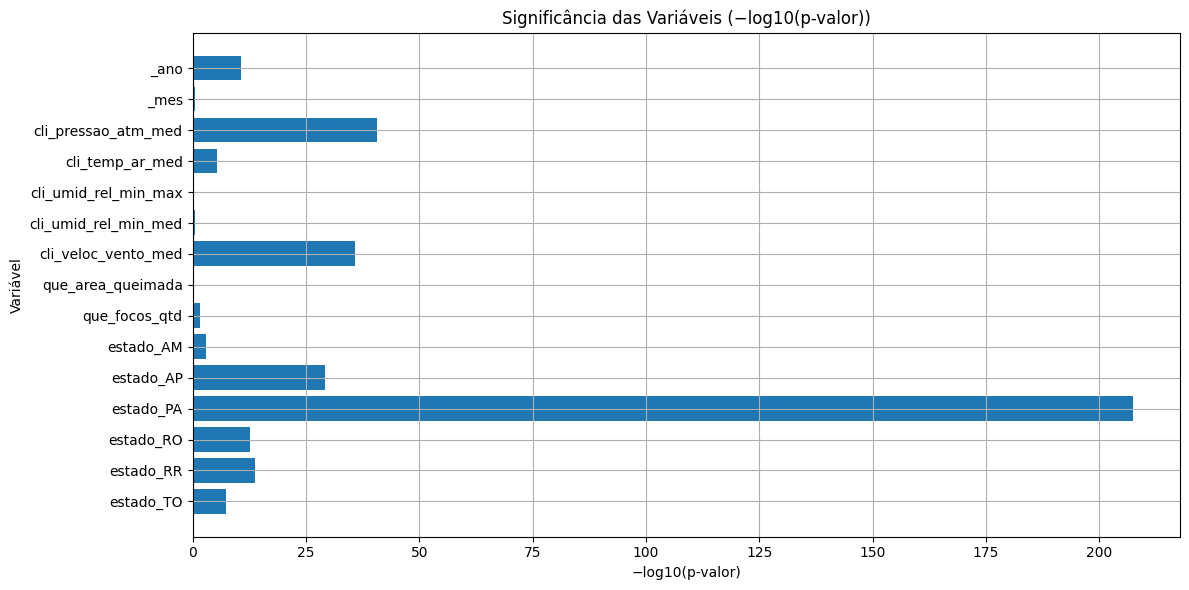
Os resultados mostram que algumas variáveis se destacam de maneira expressiva. A variável indicadora `estado\_PA` apresentou o maior coeficiente positivo, sugerindo que, estando no estado do Pará, há uma forte associação com maiores níveis de emissão de CO₂, em comparação com o estado de referência (automaticamente excluído na codificação dummy para evitar multicolinearidade). Já `estado\_AP` apresentou um coeficiente altamente negativo, indicando uma relação inversa, ou seja, estar no Amapá está associado a níveis mais baixos de emissão. Ambas as variáveis possuem p-valores próximos de zero, o que confirma sua relevância estatística.

Entre as variáveis contínuas, `cli\_veloc\_vento\_med` e `cli\_temp\_ar\_med` também se destacam. A primeira tem um coeficiente positivo elevado, o que sugere que o aumento da velocidade média do vento está associado ao aumento da emissão de CO₂, enquanto a segunda apresenta um coeficiente negativo considerável, indicando que temperaturas médias do ar mais altas tendem a reduzir a emissão, segundo a relação capturada pelo modelo. Assim como as variáveis anteriores, ambas apresentam p-valores extremamente baixos, reforçando que seus efeitos são estatisticamente significativos.

Além dessas, outras variáveis como `cli\_pressao\_atm\_med`, `\_ano` e `estado\_TO` também demonstraram significância estatística, embora com coeficientes de menor magnitude, o que indica que, embora contribuam para o modelo, seu impacto relativo é mais discreto. Por outro lado, variáveis como `\_mes`, `cli\_umid\_rel\_min\_max`, `cli\_umid\_rel\_min\_med` e `que\_area\_queimada` apresentaram p-valores altos, sugerindo que sua contribuição para a explicação da variável alvo não é estatisticamente relevante no contexto do modelo ajustado.

Portanto, com base nos coeficientes padronizados e na significância estatística, conclui-se que as variáveis `estado\_PA`, `estado\_AP`, `cli\_veloc\_vento\_med` e `cli\_temp\_ar\_med` são as que mais impactam o modelo de regressão linear múltipla, exercendo maior influência na predição da emissão de CO₂.

A significância das variáveis pode ser visualizada no gráfico abaixo, que indica o grau de confiança de que cada variável realmente influencia o modelo, sendo que quanto maior a barra, mais relevante é a variável segundo os testes estatísticos.:



5.1.3. *Interpretação dos Resultados*

(Interpretar os resultados obtidos à luz do objetivo do estudo. Relacione os achados com a questão de pesquisa e discuta se as expectativas foram atendidas. Explique as possíveis razões para os resultados observados e suas implicações.)

O modelo de Regressão Linear Múltipla aplicado neste estudo apresentou um desempenho satisfatório frente ao objetivo de prever as emissões de CO₂ na Amazônia Legal com base em variáveis ambientais e indicadores de queimadas. O coeficiente de determinação (R²) de 0,9138 indica que o modelo foi capaz de explicar mais de 91% da variabilidade observada nas emissões, o que representa um grau elevado de aderência aos dados.

Embora o desempenho geral tenha sido bom, o Erro Absoluto Médio (MAE) de 2,77 × 10⁷ toneladas revela que o modelo ainda apresenta desvios consideráveis em valores absolutos, especialmente para observações com valores muito altos de emissão. Isso sugere a presença de outliers ou de comportamentos extremos que não são totalmente capturados pela estrutura linear do modelo. O Erro Quadrático Médio (MSE) de 1,74 × 10⁵ toneladas², por outro lado, reforça que a maior parte das previsões encontra-se relativamente próxima dos valores reais.

Esses resultados atendem parcialmente às expectativas do estudo: confirmam que variáveis como área queimada, focos de incêndio e condições climáticas possuem alto poder explicativo sobre as emissões de CO₂, alinhando-se com a hipótese central de que os incêndios florestais são os principais responsáveis pelas variações de emissão atmosférica na região. No entanto, o desempenho limitado em termos de erro absoluto aponta para a necessidade de considerar, em trabalhos futuros, modelos não-lineares ou transformações na variável alvo, a fim de aumentar a precisão preditiva especialmente em cenários extremos.

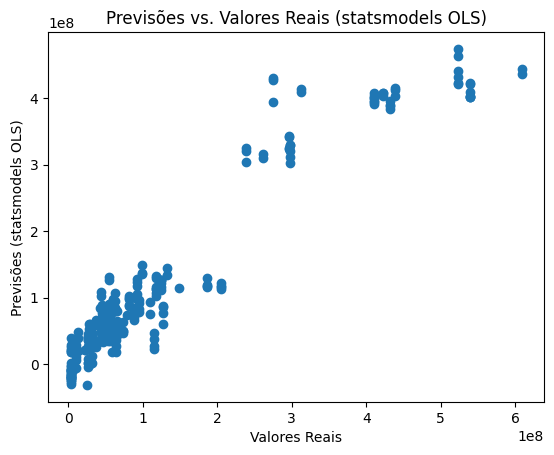
Em termos práticos, os resultados implicam que ações de controle de queimadas poderiam ter impacto direto e mensurável na redução de emissões, o que reforça a importância de políticas públicas de preservação e monitoramento ambiental na região da Amazônia Legal.

5.1.4.*Visualização dos Resultados*

Para validar os resultados obtidos com o modelo de regressão linear múltipla, foram geradas diversas visualizações que ajudam a entender a qualidade das previsões, o comportamento dos resíduos (erros) e a adequação aos pressupostos do modelo.

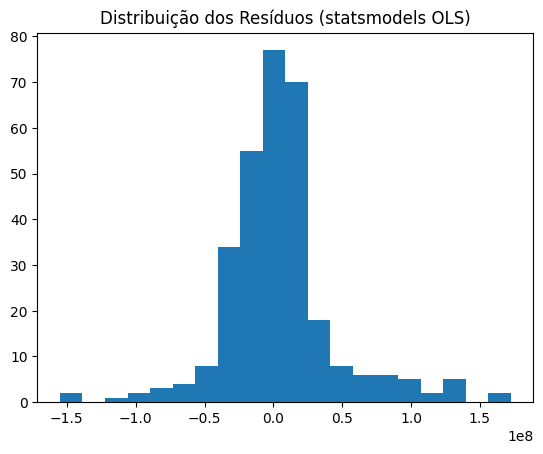
1. Previsões vs. Valores Reais

A comparação direta entre os valores reais de emissão de CO₂ e os valores previstos pelo modelo foi representada por gráficos de dispersão. Quanto mais os pontos se aproximam da diagonal imaginária (linha ideal), melhor é a capacidade do modelo de reproduzir os dados observados.



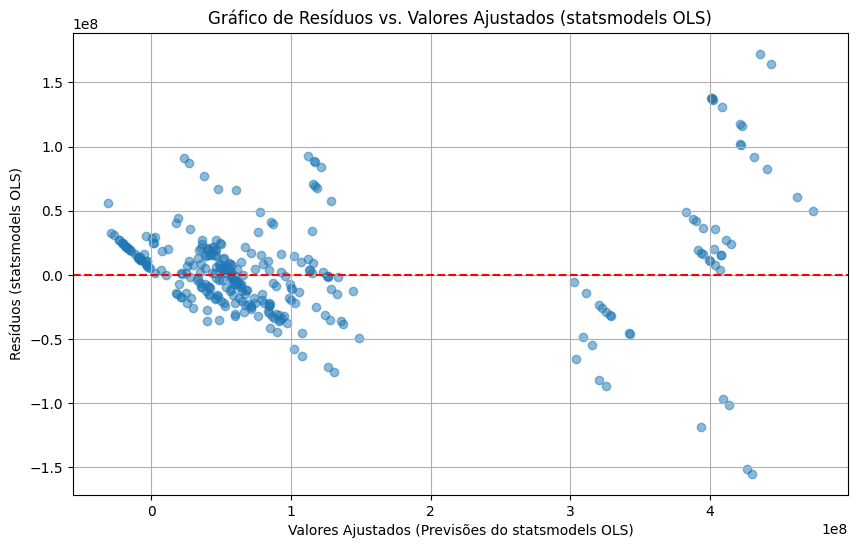
1. Distribuição dos Resíduos

Em seguida, foi analisada a distribuição dos resíduos (diferença entre o valor real e o previsto). O ideal é que esses resíduos estejam centrados em zero e distribuídos simetricamente — o que indicaria ausência de viés sistemático nas previsões. O histograma mostrou esse comportamento geral, com pequenas assimetrias nas caudas, provavelmente causadas por outliers.



1. Resíduos vs. Valores Ajustados

Por fim, foram analisados os gráficos de resíduos em função dos valores ajustados (preditos). Essa visualização é essencial para verificar a homocedasticidade, ou seja, se a variância dos erros se mantém constante em toda a faixa de valores previstos. Como pode ser visto no gráfico, os resíduos se distribuíram de forma relativamente aleatória em torno da linha horizontal (y = 0), o que sugere que esse pressuposto foi razoavelmente atendido. Contudo, uma leve dispersão nos extremos indica que o modelo pode ser sensível a valores muito altos de emissão — o que é coerente com os outliers identificados anteriormente.



Conclusão da Visualização:

As visualizações apresentadas reforçam que o modelo de regressão linear múltipla se comporta de maneira estável, com bom poder preditivo e sem grandes desvios estruturais. As análises gráficas, combinadas com as métricas de desempenho, sustentam a validade dos resultados obtidos e apontam para um modelo estatisticamente adequado à proposta do estudo.

## 5.2. Árvore de Regressão

Escolhemos testar o modelo de árvore de decisão para regressão, implementado via *DecisionTreeRegressor* da biblioteca sklearn. A escolha desse modelo se deu por vários motivos: além de ser naturalmente interpretável, permitindo a visualização das regras de decisão e a identificação das variáveis mais relevantes para a predição, as árvores de decisão apresentam flexibilidade para capturar relações não lineares e interações entre variáveis, o que pode ser uma limitação em modelos lineares tradicionais.

Antes da implementação da árvore de decisão, foi realizada uma análise utilizando a regressão linear múltipla. Embora esse modelo tenha apresentado resultados satisfatórios, observou-se que ele possui limitações na captura de relações não lineares e de interações mais complexas entre as variáveis preditoras. Por esse motivo, optou-se por testar a árvore de decisão, que, além de ser mais flexível para modelar diferentes distribuições e padrões nos dados, pode identificar interações entre variáveis e fornecer uma maior interpretabilidade ao processo de tomada de decisão. Assim, buscou-se aprimorar o desempenho do modelo e obter uma compreensão mais aprofundada dos fatores que influenciam a variável alvo.

Para a codificação, como o modelo de árvore do sklearn exige variáveis numéricas, as variáveis categóricas (como o estado) foram convertidas via codificação one-hot (pd.get\_dummies), que transforma cada categoria em uma coluna binária, permitindo que o modelo utilize essas informações.

Além disso, para o particionamento, o dataset foi dividido em conjuntos de treino e teste na proporção de 70% para treino e 30% para teste, garantindo que a avaliação do modelo seja feita em dados não vistos durante o treinamento.

A fim de garantir confiabilidade dos resultados, foi feita validação cruzada K-Fold (K=5). Isso significa que o conjunto de treino é dividido em 5 partes (folds), de modo que, em cada iteração, uma parte é usada como validação e as demais como treino. Dessa forma, permite uma avaliação mais robusta do desempenho do modelo, reduzindo o risco de overfitting e fornecendo uma estimativa mais realista da performance em dados não vistos. A função cross\_val foi implementada para automatizar esse processo, retornando os modelos treinados e as métricas de cada fold.

Após a validação cruzada, foi adotado o modelo que tenha maximizado o Coeficiente de Determinação, por ser uma métrica facilmente interpretável. Quanto às demais métricas, por serem baseadas na noção direta de distância (como MAE e MSE), seria mais difícil determinar a partir de qual faixa de erro o modelo poderia ser considerado satisfatório, especialmente sem um referencial claro para o domínio dos dados. Após a versão com maior coeficiente de determinação ser selecionada, foram realizadas previsões sobre os dados de teste. As métricas calculadas nesse conjunto refletem a capacidade do modelo de generalizar para dados realmente não vistos, simulando o desempenho em um cenário real de aplicação. Assim, a avaliação no conjunto de teste complementa a análise da validação cruzada, fornecendo uma visão mais completa e confiável sobre a performance do algoritmo.

*5.2.1. Desempenho dos Algoritmos: Árvore de Regressão*

Para a avaliação do desempenho do algoritmo foram utilizadas três métricas principais: coeficiente de determinação, erro quadrado médio e erro absoluto médio.

* Coeficiente de determinação: Mede a proporção da variância da variável alvo explicada pelo modelo. É uma métrica de fácil interpretação, sendo que valores próximos de 1 indicam que o modelo explica quase toda a variabilidade dos dados. Um valor igual a 1 representa um modelo perfeito; entre 0 e 1, indica que o modelo é melhor do que simplesmente predizer a média da variável alvo; igual a 0, significa que o modelo não melhora em relação à média; e valores negativos indicam que o modelo é pior do que predizer sempre a média. Ela foi escolhida como principal critério de seleção da versão feita pela validação cruzada por ser padrão em tarefas de regressão e por facilitar a comparação entre diferentes modelos e bases.
* Erro quadrado médio: Mede a média dos quadrados dos erros entre valores reais e preditos. É sensível a grandes erros, penalizando predições distantes dos valores reais. Utilizamos o MSE para identificar a magnitude dos erros e detectar possíveis outliers ou problemas de ajuste.
* Erro absoluto Médio: Mede a média dos valores absolutos dos erros. É menos sensível a outliers que o MSE e fornece uma noção clara do erro médio esperado em cada predição.

O resultado das métricas utilizando a validação cruzada foi o seguinte:

Média do Coeficiente de Determinação dos folds: 0.9898526950655284

Fold 0: 0.9759015584645911

Fold 1: 0.9990734586525641

Fold 2: 0.9980023973087309

Fold 3: 0.9952573054332776

Fold 4: 0.9810287554684783

Média do Erro Quadrático Médio dos folds: 184512947426747.8

Fold 0: 527535711972796.44

Fold 1: 10431920640746.988

Fold 2: 31826040605344.336

Fold 3: 103263250039995.77

Fold 4: 249507813874855.5

Média do Erro Absoluto Médio dos folds: 2500712.213191892

Fold 0: 5116202.169399955

Fold 1: 970150.0035305776

Fold 2: 996711.3369127304

Fold 3: 2411563.368359767

Fold 4: 3008934.187756429

O resultado das predições de teste (externo), utilizando o modelo com maior coeficiente de determinação:

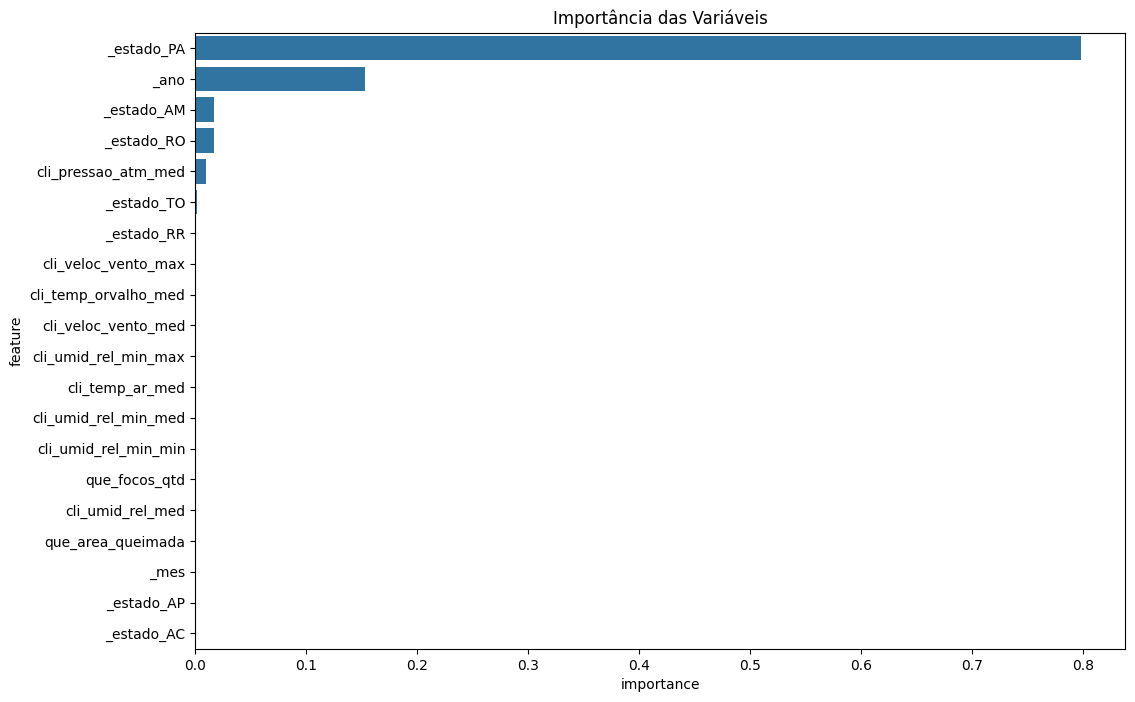
Coeficiente de Determinação: 0.9984396714545657

Erro Quadrático Médio: 31553013946686.305

Erro Absoluto Médio: 1285648.177275233

As métricas do modelo foram muito boas, com coeficiente de determinação (R²) quase igual a 1, como se o modelo fosse “perfeito”. No entanto, esse desempenho levantou suspeitas de que pode estar ocorrendo *overfitting*, ou seja, o modelo se ajusta excessivamente bem aos dados de treinamento, mas perde capacidade de generalização quando exposto a novos dados. Uma possível causa para isso é a baixa heterogeneidade do dataset, já que, como só havia dados de emissões anuais de CO₂ (variável alvo), esses valores foram imputados igualmente nos registros mensais, repetindo-se conforme o ano correspondente. Isso fez com que os dados perdessem variabilidade temporal, tornando mais fácil para o modelo aprender padrões artificiais que não necessariamente refletem a realidade das emissões mensais. Como consequência, o modelo pode parecer muito preciso nas métricas avaliadas, mas na prática não ser capaz de generalizar bem para dados reais com maior variação dentro de um mesmo ano.

## 5.2.2. Análise das Variáveis



O objetivo do projeto é estimar a quantidade de gás carbônico (CO₂) emitida em um estado, em determinado mês, com base em fatores climáticos, na quantidade de área queimada e no número de focos de incêndio. Inicialmente, esperava-se que as variáveis relacionadas às queimadas e aos fatores climáticos fossem as mais relevantes para a predição. No entanto, ao analisar o gráfico de importância das variáveis gerado pelo modelo de árvore de regressão, observou-se um resultado inesperado: variáveis como temperatura, umidade, vento, quantidade de área queimada e número de focos de queimada apresentaram importância próxima de zero.

Por outro lado, variáveis categóricas como o estado do Pará (com a maior importância, superior a 80%), o ano, o estado do Amazonas, o estado de Rondônia e, em menor grau, a pressão atmosférica média, foram as que mais influenciaram o modelo.

Esse comportamento reforça a hipótese de que o modelo está capturando padrões específicos associados aos estados e aos anos, em vez de aprender relações generalizáveis entre os fatores ambientais e de queimada com as emissões de CO₂. Isso pode ser um indício de overfitting, possivelmente causado pela baixa variabilidade dos dados e pela forma como os valores mensais de emissão de gás carbônico foram imputados com base nas emissões anuais.

## 5.2.3. Interpretação dos Resultados

Os resultados obtidos com o modelo de árvore de regressão revelaram um alto desempenho com base nas métricas avaliadas, principalmente o coeficiente de determinação, que apresentou valores próximos de 1 tanto na validação cruzada quanto no conjunto de teste. À primeira vista, isso indicaria que o modelo é capaz de prever com alta precisão a quantidade de CO₂ emitida, atendendo ao objetivo do estudo.

Entretanto, uma análise mais crítica dos resultados aponta para possíveis problemas quando se considera o objetivo principal do estudo: estimar, em determinado mês e estado, a emissão de CO₂ com base em fatores climáticos e de queimada. A expectativa inicial era de que variáveis diretamente relacionadas às queimadas e aos fatores climáticos, como temperatura, umidade, vento, quantidade de área queimada e número de focos de incêndio, tivessem alta influência sobre as emissões de CO₂. Entretanto, contrariando essa ideia, essas variáveis apresentaram importância quase nula no modelo. Por outro lado, variáveis categóricas como os estados (principalmente o Pará, com mais de 80% de importância), o ano e, em menor medida, a pressão atmosférica média, foram as que mais influenciaram as previsões do modelo. Dessa forma, há o comprometimento da capacidade do modelo de atingir o objetivo proposto, pois ele não está, de fato, utilizando os fatores ambientais e de queimada como base para suas decisões preditivas

Essa discrepância sugere que o modelo pode estar se apoiando fortemente em padrões específicos dos dados, especialmente nas categorias dos estados e dos anos, em vez de aprender relações causais ou generalizáveis entre variáveis climáticas e de queimadas e as emissões de CO₂. Isso levanta a hipótese de overfitting, ou seja, o modelo está memorizando os dados ao invés de aprender relações para generalização.

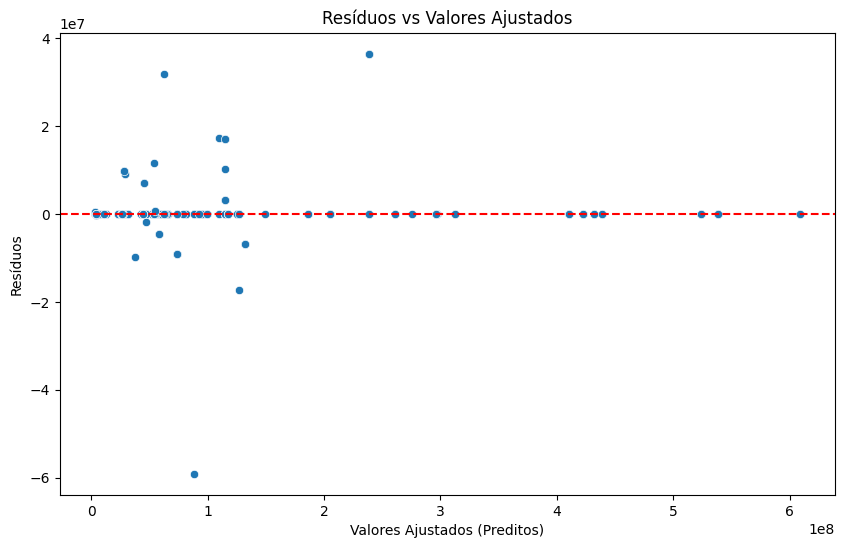
Uma provável explicação para esse comportamento está na maneira como os dados foram preparados. Como os dados de emissão de CO₂ (variável alvo) estavam disponíveis apenas em resolução anual, foi necessário imputar os mesmos valores para todos os meses de um mesmo ano em um estado. Esse procedimento, embora necessário devido a indisponibilidade de dados mensais, reduziu significativamente a variabilidade temporal.

Essa baixa variabilidade pode ter induzido o modelo a aprender que determinados estados ou anos estão diretamente associados a certos níveis de emissão de CO₂, sem de fato considerar as condições ambientais e de queimada que motivam essas emissões.

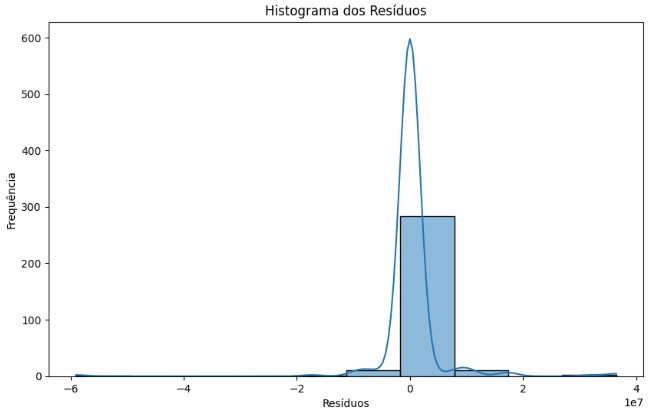
Assim, a fim de que o modelo atinja seu objetivo (estimar, mês a mês e estado por estado, a quantidade de CO₂ emitida com base em informações ambientais e de queimadas) de maneira confiável e robusta, é necessário aprimorar a base de dados, garantindo maior granularidade temporal e diversidade nos registros. Isso pode envolver a obtenção de dados ou estimativas mensais mais precisas de emissão de CO₂ e a integração com fontes alternativas de dados, como a expansão do agronegócio, que utilizam de queimadas da vegetação como meio para a abertura de novas áreas, aumentando a concentração de gás carbônico na região.

## 5.2.4. Visualização dos Resultados

A seguir, seguem análises de gráficos referentes ao modelo de árvore de regressão, apresentando visualizações claras a fim de facilitar a compreensão dos resultados.



Resíduos são definidos como a diferença entre os valores reais observados e os valores preditos pelo modelo, indicando o erro de predição. O gráfico de resíduos versus valores ajustados (ou preditos) é importante para avaliar a qualidade do ajuste do modelo de regressão. Em um modelo bem ajustado, espera-se que os resíduos estejam distribuídos de forma aleatória em torno da linha zero, sem apresentar padrões de tendências ou comportamentos repetitivos nos dados. No gráfico acima, os valores são em relação a variável alvo, quantidade de gás carbônico emitido. Observa-se que os resíduos estão bastante dispersos quando os valores preditos são baixos, com alguns pontos distantes da linha central, indicando grandes erros de predição ou outliers. Entretanto, para valores preditos mais altos, os resíduos estão mais próximos de zero, sugerindo melhor desempenho do modelo nessas situações. Essa assimetria indica que o modelo não está generalizando bem para casos de baixa emissão de CO₂, como citado anteriormente, e pode estar se ajustando demais a padrões específicos do conjunto de dados (possível overfitting já comentado).

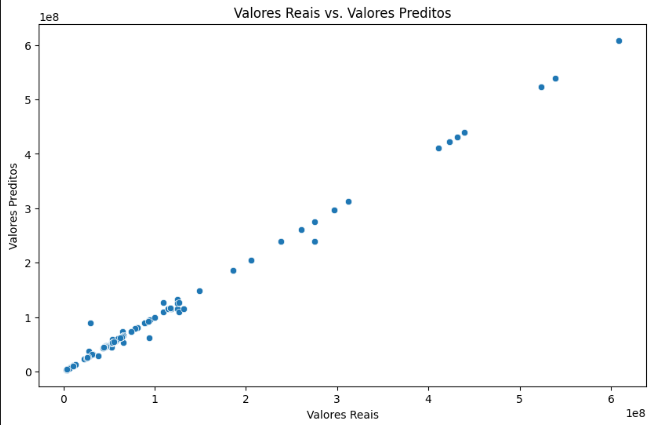


O gráfico acima mostra o histograma de resíduos do modelo, que mostra a distribuição das diferenças entre os valores reais e os valores preditos. A forma do histograma é um indicativo importante da qualidade do ajuste do modelo. Ele também complementa a análise feita anteriormente no gráfico de resíduos versus valores ajustados. Enquanto o gráfico de dispersão ajuda a verificar a presença de padrões ou curvaturas nos resíduos (indicando erros sistemáticos), o histograma nos ajuda a entender a distribuição e simetria dos erros.

Ao observar o histograma, percebemos que a maioria dos resíduos está concentrada próxima de zero, o que indica que o modelo, na maior parte das vezes, prediz valores bastante próximos dos valores reais. Essa é uma característica desejável em qualquer modelo preditivo, pois mostra que os erros não são grandes nem frequentes.

Além disso, a distribuição dos resíduos tem um formato semelhante ao de uma curva normal (distribuição gaussiana), com um pico bem definido no centro e simetria ao redor do zero. Isso reforça que os erros são, na maior parte dos casos, pequenos e distribuídos de forma aproximadamente simétrica, o que indica que o modelo não está enviesado, ou seja, ele não tende a errar sempre para mais ou para menos.

Dessa forma, a análise do histograma dos resíduos reforça os indícios de que o modelo está bem ajustado aos dados. A concentração dos resíduos em torno de zero e a forma aproximadamente simétrica da distribuição sugerem que os erros cometidos pelo modelo são pequenos, equilibrados e não sistemáticos.

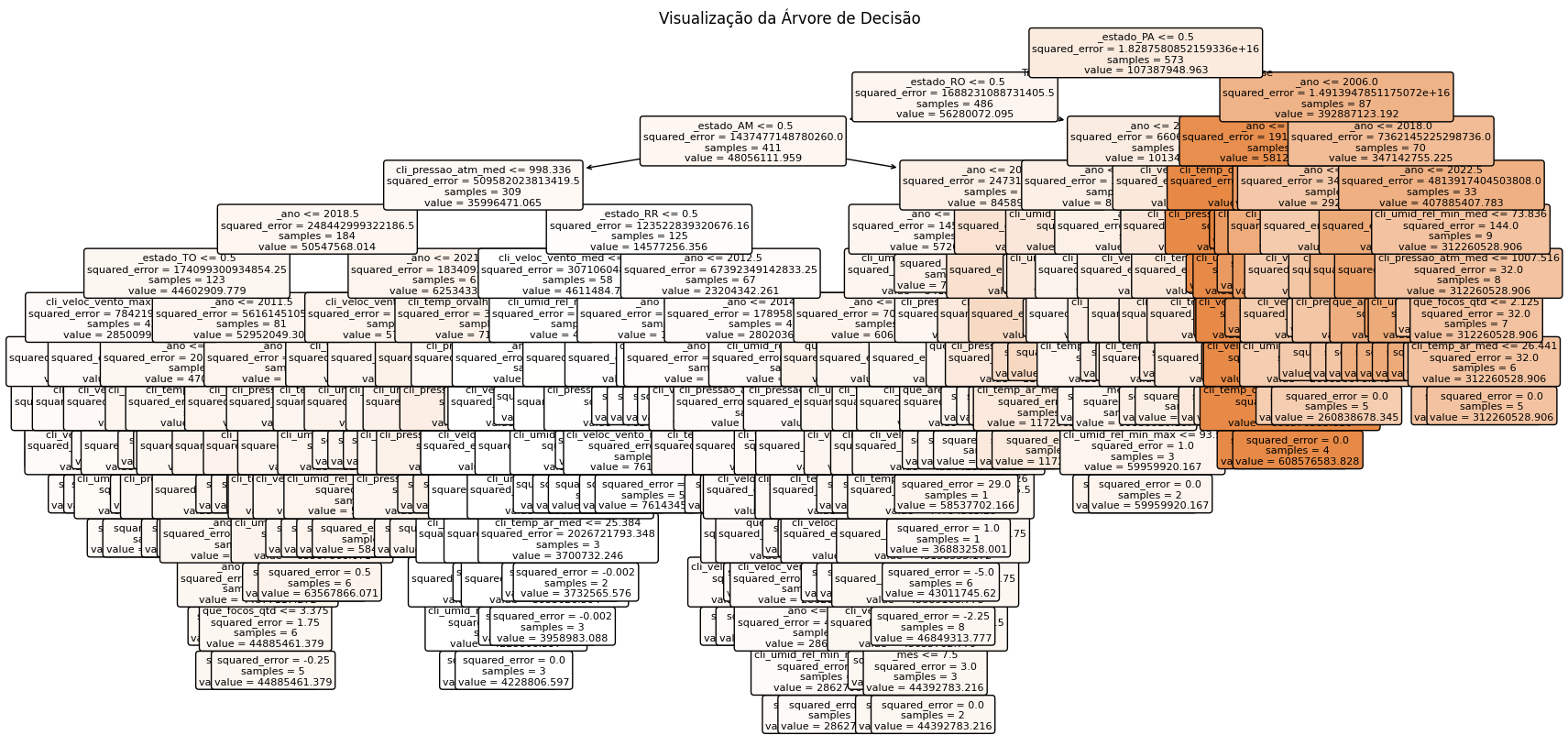


O gráfico acima representa a dispersão entre os valores reais e os valores preditos pelo modelo de árvore de regressão, sendo útil para avaliar visualmente a precisão das previsões. Cada ponto corresponde a uma observação, onde o eixo horizontal indica o valor real das emissões de CO₂ e o eixo vertical representa o valor previsto pelo modelo. Quanto mais próximos os pontos estiverem da reta y = x, melhor será o desempenho do modelo, já que isso indica que os valores preditos se aproximam dos reais.

A análise do gráfico revela que os pontos estão bem concentrados ao redor da reta y = x, com pouca dispersão, o que indica alta precisão nas previsões. Esse comportamento corrobora com o coeficiente de determinação obtido (R² = 0,9984), sugerindo que o modelo explica quase toda a variabilidade dos dados, um resultado próximo ao de um modelo perfeito.

Além disso, não são observados outliers evidentes nem padrões de curvatura que indiquem erros sistemáticos de predição. Isso sugere que, no conjunto de teste, o modelo se ajustou muito bem aos dados.

Entretanto, apesar do excelente desempenho visual, é importante interpretar esses resultados com cautela. Como discutido anteriormente, o uso de valores replicados para diferentes meses de um mesmo ano e a baixa variabilidade temporal podem ter feito com que ele memorize padrões específicos ao invés de aprender relações generalizáveis.



O gráfico apresentado é a visualização da árvore de decisão gerada pelo modelo de regressão para prever os valores de emissão de CO₂. Cada caixa representa um nó da árvore, onde ocorre uma divisão dos dados com base em uma condição sobre uma variável explicativa (por exemplo, ano, estado, temperatura, umidade, entre outras). O objetivo dessa divisão é segmentar o conjunto de dados em grupos cada vez mais homogêneos em relação à variável resposta (emissão de CO₂). Em cada nó, são mostradas informações como o erro quadrático médio (squared error), o número de amostras e o valor médio da variável resposta dentro daquele subconjunto.

A estrutura da árvore mostra que variáveis como o estado, o ano e características climáticas (pressão atmosférica, temperatura, umidade, velocidade do vento) foram fundamentais para as decisões do modelo. Por exemplo, umas das primeiras divisões são baseadas como o Amazonas, o Pará e Rondônia , o que sugere que a localização geográfica tem grande influência nas emissões. Em seguida, vemos ramificações que se aprofundam com base em variáveis climáticas e temporais, evidenciando que essas características também são altamente relevantes.

Os nós terminais (ou folhas), especialmente aqueles destacados em tons alaranjados, representam os agrupamentos finais de dados para os quais o modelo faz previsões. A intensidade da cor está associada ao valor médio de emissão de CO₂ naquele grupo, quanto mais escuro, maior o valor. Observa-se que, em algumas folhas, o erro quadrático é zero, o que significa que, para aquele subconjunto, o modelo previu exatamente os valores reais, o que pode ocorrer quando há repetição ou pouca variabilidade nos dados daquele grupo.

É interessante destacar também a profundidade da árvore, o que indica que o modelo foi bastante ajustado aos dados (o que chamamos de overfitting em alguns casos). Embora o ajuste aparente seja bom, como vimos nas análises anteriores (gráfico de valores reais vs. preditos e resíduos), árvores muito profundas podem generalizar mal em novos dados, por captarem ruídos ou padrões muito específicos do conjunto de treino.

# Referências

ECOA PUC-RIO. **AMazonizAR**. *Homepage* da instituição, 2023. Disponível em:<https://instituto.ecoa.puc-rio.br/amazonizar/>. Acesso em: 29 mar. 2025.

IBGE. **IBGE atualiza mapa da Amazônia Legal**. Agência de Notícias IBGE, 16 jul. 2020. Disponível em:<https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/28089-ibge-atualiza-mapa-da-amazonia-legal>. Acesso em: 4 abr. 2025.

INSTITUTO NACIONAL DE METEOROLOGIA (INMET). *Dados históricos*. Brasília: INMET, [s.d.]. Disponível em: <https://portal.inmet.gov.br/dadoshistoricos>. Acesso em: 2 maio 2025.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). *Programa Queimadas*. São José dos Campos: INPE, [s.d.]. Disponível em: <https://terrabrasilis.dpi.inpe.br/queimadas>. Acesso em: 2 maio 2025.

MAPBIOMAS. *Monitor do Fogo*. São Paulo: Projeto MapBiomas, [s.d.]. Disponível em: <https://plataforma.brasil.mapbiomas.org/monitor-do-fogo>. Acesso em: 2 maio 2025.

OBSERVATÓRIO DO CLIMA. **Sistema de Estimativas de Emissões de Gases de Efeito Estufa (SEEG)**. Disponível em:<https://plataforma.seeg.eco.br/>. Acesso em: 29 mar. 2025.