UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

Andressa Luana Santana de Lima

Aprendizado de Máquina Não Supervisionado Aplicado na Dinâmica de Preços de Combustíveis no Brasil

Recife

Agosto de 2025

Andressa Luana Santana de Lima

Aprendizado de Máquina Não Supervisionado Aplicado na Dinâmica de Preços de Combustíveis no Brasil

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE
Departamento de Estatística e Informática
Curso de Bacharelado em Sistemas de Informação

Orientador: Roberta Macedo Marques Gouveia

Recife
Agosto de 2025

ANDRESSA LUANA SANTANA DE LIMA

APRENDIZADO DE MÁQUINA NÃO SUPERVISIONADO APLICADO NA DINÂMICA DE PREÇOS DE COMBUSTÍVEIS NO BRASIL

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada em: 05 de Agosto de 2025.

BANCA EXAMINADORA

Roberta Macêdo Marques Gouveia (Orientadora)
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Elizabeth Regina Tschá
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Aprendizado de Máquina Não Supervisionado Aplicado na Dinâmica de Preços de Combustíveis no Brasil

Andressa Luana Santana de Lima[1], **Roberta Macedo Marques Gouveia**[2]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco
Rua Dom Manuel de Medeiros, s/n, - CEP: 52171-900 – Recife – PE – Brasil

andressa.santana@ufrpe.br, roberta.gouveia@ufrpe.br

Abstract. *This study presents an exploratory and clustering analysis of public data from the Agência Nacional do Petróleo (ANP) on fuel prices in 2024. Based on numerical variables aggregated by region and by product, the K-means algorithm was applied to identify behavioral patterns in the market. The selected variables aimed to represent aspects such as average price levels, seasonal variations, record volume, and reseller distribution. The results revealed structural differences between regions and fuel types, highlighting the sector's heterogeneity. The study reinforces the importance of using clustering techniques to explore relevant patterns in the fuel market.*

Resumo. *Este trabalho realiza uma análise exploratória e de clusterização dos dados públicos da Agência Nacional do Petróleo (ANP) para os preços de combustíveis no Brasil em 2024. A partir de variáveis numéricas agregadas por região e por produto, foi aplicado o algoritmo K-means para identificar padrões de comportamento no mercado. As variáveis selecionadas buscaram representar aspectos como níveis médios de preço, variações sazonais, volume de registros e distribuição de revendas. Os resultados apontaram diferenças estruturais entre regiões e entre combustíveis, evidenciando a heterogeneidade do setor. O estudo evidencia a importância do uso de técnicas de agrupamento para explorar padrões relevantes no mercado de combustíveis.*

1. Introdução

A análise de grandes volumes de dados tem se tornado cada vez mais relevante em diversas áreas, especialmente no monitoramento do comportamento de mercado e na definição de preços. No setor de combustíveis, onde a variabilidade regional, os custos logísticos e a carga tributária influenciam diretamente os valores praticados, compreender esses padrões pode contribuir com as investigações acadêmicas voltadas à dinâmica de precificação do setor.

A flutuação dos preços dos combustíveis no Brasil tem sido objeto de debate recorrente, dado seu impacto direto na inflação, na competitividade de setores produtivos e na economia. Essa volatilidade ocorre tanto por fatores externos, como a variação do barril de petróleo e da taxa de câmbio quanto por decisões internas relacionadas à política de preços, subsídios e estrutura tributária. A precificação dos combustíveis no Brasil está diretamente ligada à cotação internacional do petróleo e à taxa de câmbio, uma vez que o petróleo é uma *commodity* cotada em dólares. Dessa forma, elevações no valor do barril ou na taxa de câmbio tendem a pressionar os preços internos dos combustíveis (DELGADO; GAUTO, 2021). Além desses fatores, a guerra entre Rússia e Ucrânia, iniciada

em fevereiro de 2022, provocou impactos expressivos nos mercados globais de energia, ao interromper cadeias de suprimento de petróleo e gás, gerando volatilidade nos preços. Esse cenário contribuiu para elevações significativas nos preços da energia em escala mundial, afetando particularmente países dependentes de importações de combustíveis fósseis (ALSAIFI, 2023).

Eventos internacionais, como os choques do petróleo em 1973 e 1979, historicamente evidenciaram o impacto direto de crises geopolíticas sobre a inflação e a estabilidade macroeconômica de diversos países, inclusive o Brasil (STASZCZAK, 2019). Além disso, a literatura recente aponta que a ausência de estabilidade regulatória e a presença de interferências políticas e redes de interesse na formulação e implementação de políticas energéticas favorecem respostas fragmentadas entre as esferas do governo, ampliando disparidades de preços entre as regiões (DE ANDRADE e RODRIGUES, 2024).

Neste trabalho, a técnica de clusterização *K-means* é proposta para identificar perfis regionais e segmentações por tipo de combustível a partir da base de dados da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), com recorte do ano de 2024. A motivação para o desenvolvimento deste estudo está relacionada à importância dos combustíveis no contexto nacional e às variações significativas de preços observadas entre regiões e produtos. Considerando a dimensão continental do Brasil, as diferenças na logística de distribuição e a diversidade tributária por localidade, compreender como esses fatores se refletem nos dados é fundamental para investigar a dinâmica de precificação sob uma perspectiva analítica e segmentada. A partir dessa abordagem, busca-se identificar padrões que contribuam para o entendimento do mercado de combustíveis.

A metodologia adotada envolveu etapas de coleta, transformação e aplicação do algoritmo *K-means*, com o número ideal de grupos definido pelo método do cotovelo. Em seguida foi realizada a interpretação dos *clusters* com o auxílio da Análise de Componentes Principais (*PCA*).

Espera-se que os resultados obtidos contribuam para o entendimento da dinâmica de preços e da estrutura de comercialização dos combustíveis no Brasil, destacando as disparidades regionais e comportamentos específicos por tipo de produto.

Este artigo está dividido em sete seções: 1. Introdução, 2. Referencial teórico, 3. Trabalhos relacionados, 4. Metodologia, 5. Análise Exploratória dos dados 6. Resultados e 7. Conclusões.

2. Referencial teórico

O Aprendizado de Máquina, ou *Machine Learning* (ML), é uma área da Ciência da Computação, mais especificamente, é uma sub área da Inteligência Artificial voltada para o desenvolvimento de algoritmos que aprendem através dos dados de entrada e geram conhecimento automaticamente (MENEZES et al., 2024). Com o avanço da tecnologia, o ML tem demostrado ser uma ferramenta robusta para a análise de grandes volumes de dados e identificação de padrões complexos, sendo amplamente aplicada em diversas áreas do conhecimento para fazer predições, agrupamentos ou tomar decisões orientadas por dados.

As três principais técnicas de aprendizagem em ML, conforme abordada por Menezes et al. (2024) são: a supervisionada, a não supervisionada e por reforço. A apren-

dizagem supervisionada é declarada como um conjunto de dados de entrada e saída com o objetivo de treinar o algoritmo para encontrar uma regra que identifique padrões e realize previsões ou classifique novos dados. Esse tipo de abordagem é aplicado para as categorias de classificação e regressão. Por outro lado, a não supervisionada, os dados de entrada e saída não são informados para encontrar padrões, ou seja, eles não possuem rótulos. Neste caso, o algoritmo precisa descobrir grupos de forma natural. Um exemplo são os modelos de agrupamento, como é o caso do *K-means* que iremos aprofundar neste estudo nas próximas seções. Por fim, o aprendizado por reforço se baseia em um modelo que aprende por meio de decisões sequenciais e também de seus erros, após várias tentativas, para chegar em uma solução.

2.1. Clusterização dos dados

De acordo com Sinaga and Yang (2020), a clusterização é uma ferramenta útil na área da ciência de dados, pois permite identificar estruturas de agrupamento em um conjunto de dados com base em similaridades internas e diferenças externas entre os grupos. É uma abordagem exploratória que possibilita extrair padrões mesmo na ausência de rótulos ou classificações previamente definidas.

De acordo com a classificação apresentada pelos autores, os métodos de clusterização podem ser divididos em duas categorias principais: baseados em modelo e os não paramétricos. A primeira representa os métodos baseados em modelos estatísticos, onde consideram que os dados foram gerados por uma combinação de distribuições matemáticas. Um exemplo é o algoritmos Expectação-Maximização (EM) que é utilizado para estimar os agrupamentos com base nas probabilidades. A segunda abordagem, não envolve a aplicação dos modelos estatísticos, pois usam medidas de similaridade ou semelhança para a formação os grupos. Nesta categoria, há dois subtipos metodológicos: o método hierárquico que foi um dos primeiros utilizados principalmente por cientistas sociais e biólogos e o método particional que divide os grupos de dados de forma finita por um valor, como é o exemplo do *K-means*

Na próxima subseção, apresenta com mais detalhes o funcionamento do *K-means* e o uso do método do cotovelo como estratégia adotada para definição do número ideal de agrupamentos.

2.1.1. Algoritmo de Clusterização K-means

O algoritmo *K-means* é uma técnica clássica e bastante aplicada para análise clusterizações em bases de dados que tem o objetivo de dividir um conjunto de dados em k grupos distintos, buscando minimizar a distância entre os pontos e os centróides dos *clusters*. O processo é feito de forma gradual: os dados são atribuídos ao centróide mais próximo, os centróides são atualizados com base na média dos dados atribuídos a eles, e essa etapa ocorre de forma periódica até que não haja mais alterações nas atribuições.

Essa abordagem é amplamente utilizada por sua simplicidade e desempenho, sendo aplicada em reconhecimento de padrões, processamento de dados e mineração de dados, conforme descrito por (PATIL et al., 2024). Neste projeto, o algoritmo foi aplicado para reconhecer os padrões de comportamento dos combustíveis no Brasil durante o ano de 2024.

Como citado anteriormente, é necessário definir o valor de k, que representa a quantidade de *clusters* a serem formados após a aplicação do algoritmo. Uma das formas mais utilizadas para essa definição é através do método chamado Elbow (cotovelo, em português), ao invés de ser escolhido manualmente de forma aleatória. Essa técnica baseia-se na análise gráfica da variação da Soma dos Erros Quadráticos (SSE, do inglês *Sum of Squared Errors*), à medida que diferentes valores de k são testados, em uma iteração, por exemplo. O objetivo é observar em que ponto o acréscimo de novos *clusters* deixam de trazer uma redução significativa do erro, formando uma curva com um formato semelhante a um cotovelo. Esse ponto de inflexão representa o valor ideal de k.

De acordo com Wala et al. (2024), as etapas para aplicação do método Elbow incluem: definir um intervalo de valores para k (por exemplo, de 1 a 10), aplicar o algoritmo *K-means* para cada valor, calcular o *SSE* correspondente, plotar os resultados em um gráfico e identificar visualmente o ponto onde ocorre a maior queda antes da curva se estabilizar. Neste trabalho, essas etapas foram seguidas, sendo mais detalhado posteriormente na seção de resultados.

Esta mesma autora indica que o *SSE* representa o grau de coesão dos dados dentro de cada *cluster*, sendo calculado pela soma das distâncias quadráticas entre os pontos e seus respectivos centróides, como mostrado na Equação 1:

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} (x_i - c_j)^2 \tag{1}$$

onde k é o número total de *clusters*, n é o número de pontos atribuídos ao *cluster* j, x_i representa um ponto de dado, e c_j é o centróide do *cluster* correspondente.

Um *SSE* menor indica *clusters* mais coesos. No entanto, aumentar o número de *clusters* além do necessário tende a reduzir o *SSE*, gerando segmentações mais significativas. Por isso, o método *Elbow* é utilizado para encontrar um equilíbrio e indica um valor mais ideal para a aplicação do algoritmo.

Alguns algoritmos de aprendizado de máquina, como é o caso do *K-means*, utilizam apenas dados numéricos. Dessa forma, a conversão de variáveis categóricas em valores numéricos é um processo fundamental a ser realizado antes da etapa do pré-processamento, fazendo com que os modelos identifiquem padrões e gerem previsões de forma adequada.

As estratégias mais comuns para fazer essa conversão são: *Label Encoding* e o *One-Hot Encoding*. O *Label Encoding* atribui um número inteiro a cada categoria. Por exemplo, para a variável ”Modalidade de Pagamento” que possui os valores “Cartão”, “Dinheiro”, “Pix” e ”Boleto”, o modelo poderia representar Cartão = 0, Dinheiro = 1, Pix = 2 e Boleto = 3. No entanto, esse método pode causar o entendimento de ordem, o que pode afetar o desempenho do algoritmo em contextos em que as categorias não dependem da ordem. Por outro lado, o *One-Hot Encoding* evita esse problema, pois essa estratégia cria uma nova coluna para cada categoria, preenchendo com 1 quando o dado pertence a categoria e com 0 quando não pertence. Por exemplo, reutilizando a variável anterior “Modalidade de Pagamento” e os seus valores, citados no paragrafo anterior, seriam transformados em quatro colunas respectivamente. Se uma transação foi paga

em Pix, apenas a coluna “Pix” recebe valor 1, enquanto as demais recebem 0. Como esse método não indica a relação ordinal entre as categorias, ele é mais adequado para o tipo de dado utilizado neste estudo. Portanto, foi utilizado o *One-Hot Encoding* para transformar os atributos categóricos em numéricos que será detalhado posteriormente.

O objetivo do uso desta estratégia foi evitar que a variável categórica recebessem pesos de ordenação, como ocorreria com o método *Label Encoding*, por exemplo: ao transformar as categorias das cores verde, vermelho e amarelo em [0,1,2]. O *One-Hot Encoding* cria uma coluna binária para representar as categorias e atribui o valor 1 para a variável que representa a categoria e as demais recebem o 0. Seguindo o exemplo anterior, a categoria vermelho ficaria: [0,1,0], indicando verde, vermelho e amarelo, respectivamente.

3. Trabalhos Relacionados

Nesta seção, são apresentados trabalhos científicos que utilizaram a base de dados da ANP, relacionados ao tema deste artigo.

No artigo, Quintino et al. (2022), os autores analisaram a tendência temporal e a correlação dos preços relativos entre o bioetanol e gasolina nas 15 principais capitais brasileiras, utilizando técnicas como *Detrended Fluctuation Analysis* (DFA) e *Detrended Cross-Correlation Analysis* (DCCA). O estudo analisou um período de 2004 a 2020, identificando padrões distintos por região e evidenciando que o mercado de combustíveis no Brasil possui uma forte característica regional. A base de dados da ANP foi essencial para avaliar a eficiência do mercado, onde os seus resultados são relevantes para criação de políticas públicas e estratégias de distribuição. No entanto, os autores não exploraram agrupamentos entre regiões, não aplicaram técnicas de segmentação como algoritmos de clusterizações e não realizaram comparações com outros tipos de combustíveis, além do bioetanol e gasolina.

Utilizando a mesma base de dados da ANP, Mosquéra et al. (2024) investigaram a dinâmica entre o preço relativo do etanol e da gasolina e como impactava na escolha do consumidor em diferentes estados brasileiros, buscando identificar os padrões de consumo. O estudo classificou cada estado em três tipos de faixas de acordo com a competitividade do etanol, considerando os aspectos como infraestrutura e da logística na determinação dos preços em nível regional. Através da análise dos parâmetros de preço, elasticidades cruzadas e participação de veículos que funcionam com diferentes tipos de combustíveis, chamados de *flex fuel*, os autores conseguiram identificar padrões de consumo e sugerir cenários de incentivo à adoção de combustíveis renováveis. Apesar de oferecerem uma contribuição significativa para a compreensão da competitividade entre combustíveis, os autores concentraram sua análise apenas na relação entre o etanol e a gasolina. Essa abordagem não explorou a diversidade de combustíveis disponíveis no mercado e nem a variação das características de mercado por região do país.

Por fim, no estudo do Dos Santos Barcellos (2017), desenvolveu um protótipo de uma aplicação, com base na análise da série histórica de preços de combustíveis, utilizando dados obtidos da ANP para visualizar a variação dos preços por meio de gráficos interativos. O autor identificou que os consumidores brasileiros enfrentavam dificuldades para acompanhar a flutuação dos preços ao longo do tempo e em diferentes localidades. A validação do protótipo foi realizada através de um formulário com usuários, que avali-

aram a utilidade e a usabilidade da aplicação desenvolvida. Embora o trabalho do autor tenha relevância na usabilidade das informações para os consumidores dos combustíveis, não foram realizada análises estatísticas para investigar os padrões e comportamento dos dados.

Este artigo, apresenta uma abordagem de análise exploratória e de clusterização utilizando diversas variáveis numéricas e categóricas derivadas dos dados públicos disponibilizados pela ANP. A proposta é identificar padrões entre os combustíveis e entre as regiões do Brasil, aplicando técnicas de agrupamento para indicar as semelhanças e as diferenças. Dessa forma, este estudo compartilha uma nova forma de segmentação utilizando os dados da ANP, explorando padrões não evidenciados nos estudos anteriores.

4. Metodologia

A metodologia utilizada para o desenvolvimento deste trabalho foi dividida em quatro etapas principais: coleta dos dados, transformação, aplicação do *K-means* e estudo dos resultados. A primeira etapa consistiu na coleta dos dados disponibilizados pela ANP referente ao ano de 2024. Em seguida, foi realizada a etapa de transformação dos dados, com o objetivo de organizar e estruturar os dados para as análises posteriores, incluindo a criação de nova colunas. Antes da aplicação do algoritmo, foi realizada uma análise exploratória dos dados, com o objetivo de compreender o comportamento inicial dos dados, identificar distribuições relevantes, detectar possíveis *outliers* e reconhecer particularidades nas variáveis analisadas. Essa etapa está detalhada em uma seção separada. Na sequência, foi aplicado o método do cotovelo (*Elbow Method*) para definir o número ideal de agrupamentos, seguido da implementação do algoritmo de clusterização *K-means*. Para facilitar a visualização dos resultados, foi utilizada a técnica de Análise de Componentes Principais (PCA), permitindo a representação bidimensional dos *clusters*. Por fim, a última etapa foi dedicada ao estudo e interpretação dos agrupamentos gerados, com o objetivo de coletar *insights* relevantes sobre os padrões de preços de combustíveis no Brasil.

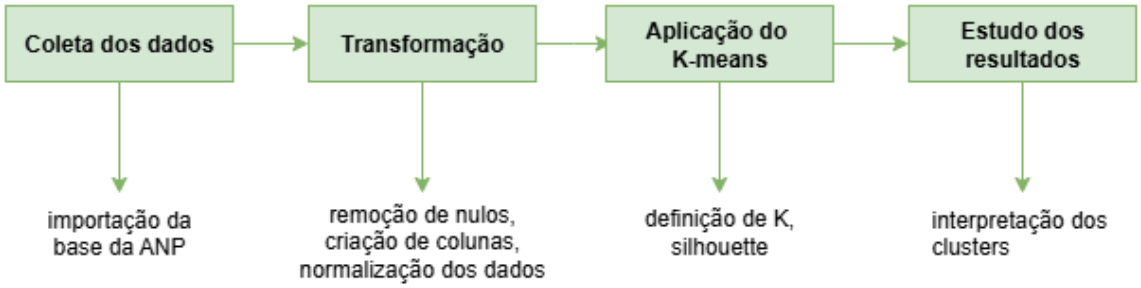

Figura 1. Fluxo das etapas da metodologia.
Fonte: A autora (2025).

Durante a metodologia e a análise exploratória dos dados, foi utilizada a linguagem de programação Python na versão 3.11.12, no ambiente colaborativo Google Colab. As bibliotecas utilizadas para o desenvolvimento do projeto foram: Pandas para manipulação de dados, o NumPy para operações numéricas, ydata-profiling para identificar os *outliers*, o Scikit-Learn para pré-processamento e aplicação de algoritmos de

aprendizado de máquina, o Matplotlib e o Seaborn, ambos, para visualização gráfica dos dados.

O processo aplicado nesta seção está presente na Figura 1 que descreve essas etapas.

4.1. Coleta de dados e transformação de dados

Esse projeto iniciou com a coleta dos dados, disponibilizados pela ANP, referentes aos preços médios dos combustíveis no Brasil, pelo recorte do ano de 2024. A fonte de dados utilizada possui 16 colunas e 898536 linhas de registros como tipo de combustível, unidade da federação, município, bandeira do posto, valor de venda, data de referência, entre outras. A partir desses dados, as análises e modelagens desenvolvidas nas etapas posteriores deste estudo foram realizadas.

Em seguida, foi executada a transformação, etapa em que ocorre o processo de limpeza de dados, remoção de linhas duplicadas, criação e padronização de variáveis para preparar a base para as análises e modelagens.

Além isso, foram removidas algumas variáveis consideradas desnecessárias para o estudo deste projeto, a fim de diminuir a dimensionalidade da tabela e manter apenas as informações relevantes, tais como: Complemento, Valor de Compra, Bairro, Nome da Rua, Numero Rua e Cep. As linhas de registros duplicadas foram excluídas, foram cerca de 20 linhas com esse comportamento e foi mantido apenas a primeira ocorrência do dado. Nessa etapa também foi feita a renomeação dos campos seguindo o padrão snake_case, onde os nomes das colunas são escritos em minusculo e se a palavra for composta é acrescentado o *underline* para compor o nome da coluna. Exemplo: o atributo se chamava Valor de venda, passou a ser valor_venda e assim por diante.

Alguns atributos presentes na base de dados original continham siglas, mas dessa forma não era amigavél para compreender a informação. Para resolver essa questão, foi aplicado um procedimento "de-para", no qual os valores originais foram mapeados para nomenclaturas mais descritivas e legíveis. Por exemplo: a coluna regiao, que inicialmente apresentava siglas como NE, foram padronizados para valores descritivos, no caso em especifico, para Nordeste. Essa padronização visou facilitar a interpretação e evitar ambiguidade durante a análise e a visualização dos dados. Nessa etapa também ocorreram conversões de *low case*, formatação de data e de valores. Após os passos acima, foram criadas várias colunas a partir das variáveis originais para possibilitar análises mais refinadas e otimizar a modelagem. A maioria das variáveis adicionadas contribuiu com diferentes níveis de granularidade, principalmente em relação as dimensões de tempo (safra) e localização (município, estado e regiao), além da categorização por produto. As principais variáveis criadas são:

- safra_coleta: coluna criada a partir da data de coleta no formato YYYYMM (ano e mês), permitindo análises mensais agregadas.
- preco_medio_produto_*: representa a média de preço por produto considerando diferentes granularidades: total, por safra, por município, estado e região.
- var_preco_media_produto_*: são colunas que indicam a variação do preço de venda em relação à média nas mesmas granularidades anteriores.
- qnt_coletas_*: número de coletas realizadas por produto em diferentes recortes geográficos e temporais.

- qnt_produtos_revenda_*: quantidade de produtos distintos por revenda, considerando as visões por safra, município, estado e região.
- qnt_revendas_*: número de revendas por safra, produto e localização geográfica.
- preco_relativo_*: preço de venda relativo à média (igual a 1 significa igual à média; maior que 1, acima da média; abaixo de 1, abaixo da média), útil para análises comparativas.

As colunas criadas foram fundamentais tanto para as análises exploratórias quanto para a definição das variáveis de entrada nos algoritmos de clusterização. Elas permitiram organizar de forma mais estruturada o comportamento dos preços de combustíveis no Brasil ao longo de 2024, o que facilitou para a identificação de padrões relevantes nas visões de região e produto.

4.2. Aplicação do K-means

Posteriormente, ocorreu a fase de implementação do *K-means* para segmentar o conjunto de dados em grupos com características similares. A base final de dados para a aplicação deste algoritmo contém 42 colunas e 898526 linhas de registos, onde cada linha representa um registro único de coleta de preço de revenda por produto, safra, município, estado e região, já enriquecido com variáveis derivadas como preço médio, variação de preço, volume de coletas e quantidade de revendas. A granularidade combina dimensões temporais e geográficas, sem duplicidade de registros. Essas variáveis podem ser visualizadas no Apêndice A: Dicionário de Dados.

A aplicação do *K-means* requer que as colunas sejam compostas por dados numéricos, sendo necessário transformar as variáveis categóricas em variáveis numéricas. Na base construída, além das variáveis numéricas derivadas, também há variáveis categóricas como município, estado, região, produto, bandeira e revenda. No entanto, para a implementação do algoritimo, apenas as colunas de região e produto foram escolhidas e utilizadas de forma separada em cada análise. Ambas são representações mais agregadas e estão relacionadas ao objetivo deste estudo que é identificar padrões na visão regional e por tipo de produto. Para essa conversão, adotou-se o método de *One-Hot Encoding*, utilizando a classe *OneHotEncoder()* da biblioteca *scikit-learn* (versão 1.6.1).

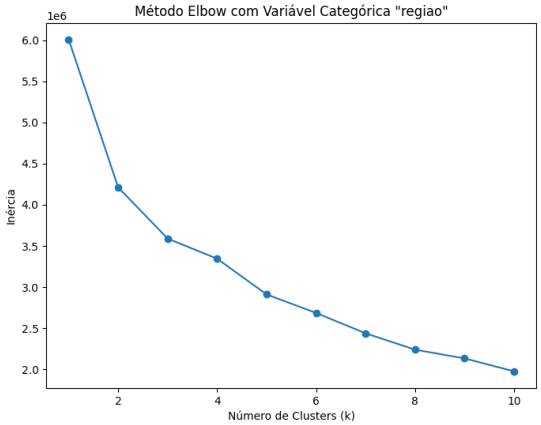

Gráfico 1. Método do cotovelo para a visão de região.
Fonte: A autora (2025).

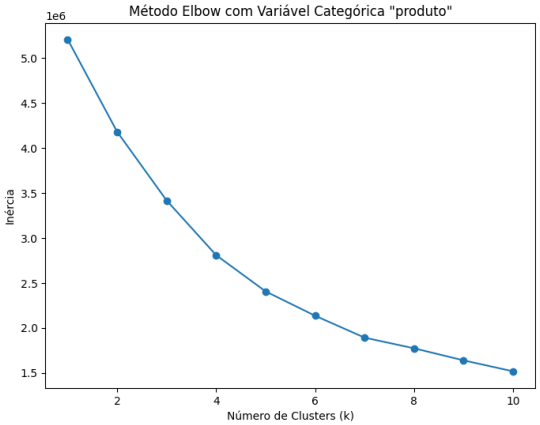

Gráfico 2. Método do cotovelo para a visão de produto.
Fonte: A autora (2025).

A fim de evitar distorções nas distâncias calculadas pelo algoritmo, as variáveis numéricas foram normalizadas por meio do *StandardScaler()*, que transforma cada variável para que tenha média zero e desvio padrão igual a um. Esse tratamento foi essencial para colocar as variáveis na mesma escala e evitar que aquelas com valores originalmente mais altos tivessem um peso maior na formação dos grupos.

O *K-means* precisa de um valor de entrada definido previamente para k, onde foi abordado na seção de Referencial Teórico deste trabalho. Para isso, foi aplicado o método do cotovelo, também conhecido como *elbow*, que compara a variação da soma dos erros quadráticos internos (inertia) para diferentes valores de k. Neste estudo o melhor número de k foram: k = 3 para a visão de região e k = 4 para a de produto. Nos Gráficos 1 e 2, estão representados o resultado dos gráfico *elbow* para cada uma dessas análises.

Após a etapa de clusterização com o algoritmo *K-means*, os resultados foram incluídos à base de dados com a criação de uma coluna chamada *cluster*, que identifica qual grupo cada registro pertence. Com essa informação foi possível realizar a fase de exploração dos *clusters* formados, analisando as semelhanças e diferenças entre eles. A análise teve o objetivo de entender como as variáveis de preço, quantidade, produto e região se agruparam, buscando identificar padrões e *insights* relevantes sobre o mercado de combustíveis que são abordados na seção de Resultados deste estudo.

Para avaliar a qualidade dos agrupamentos gerados pelo algoritmo *K-means*, foi aplicado o método da silhueta, conforme descrito por (SINAGA and YANG, 2020). Essa abordagem permite validar se cada ponto de dado realmente pertence ao seu *cluster*, considerando tanto separação em relação aos outros *clusters* quanto à coesão interna no grupo.

O cálculo do coeficiente de silhueta $s(i)$ possui a fórmula:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

em que:

- $a(i)$ é a distância média do ponto i para os demais pontos do mesmo *cluster*;
- $b(i)$ é a menor distância média do ponto i para os pontos de outros *clusters*.

O resultado de $s(i)$ varia entre o intervalo de -1 e 1. Os valores próximos de 1, indicam que o ponto está bem agrupado. Os valores próximos de 0, representa que o ponto está próximo à borda dos *clusters* e negativos indicam que o ponto pode ter sido inserido no *cluster* errado. A média dos valores de $s(i)$ para todos os pontos permite identificar uma medida global da qualidade da clusterização.

Por fim, foi conduzido um estudo dos resultados obtidos para descrever as principais características dos agrupamentos e identificar padrões relevantes para a compreensão da dinâmica dos preços e perfis de comercialização dos combustíveis no Brasil. Essa etapa consistiu na análise dos *clusters* gerados, onde as suas interpretações estão apresentados na seção 6.

5. Análise Exploratória dos Dados

Antes da aplicação do algoritmo de clusterização, foi realizada uma análise exploratória com o objetivo de compreender a estrutura da base de dados, avaliar a representatividade

das observações por localidade e produto, além de identificar padrões iniciais de comportamento dos preços dos combustíveis no Brasil ao longo de 2024. Essa etapa auxiliou a escolha das variáveis mais adequadas para o modelo de agrupamento e contribuiu para uma melhor interpretação dos resultados.

5.1. Distribuição das Observações

Inicialmente, foram avaliadas as distribuições percentuais das coletas por estado, região e tipo de produto.

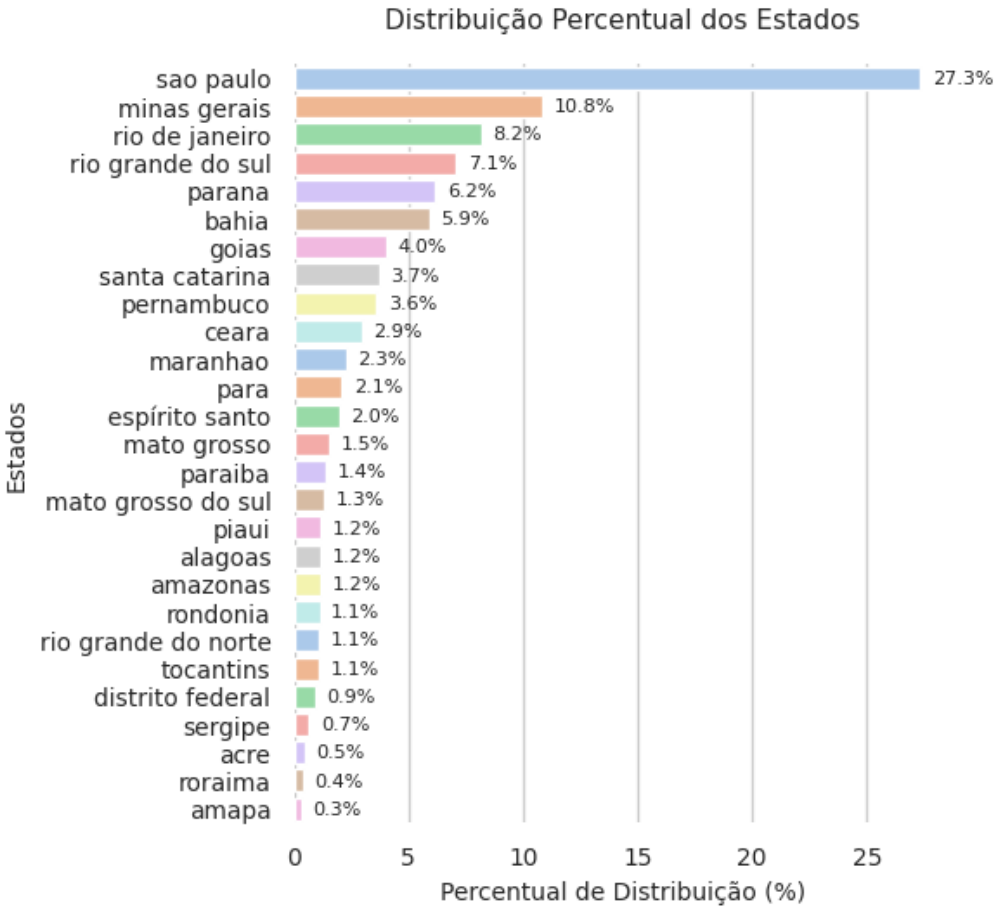

Gráfico 3. Distribuição percentual das observações por estado.
Fonte: A autora (2025).

O Gráfico 3 acima, exibe a distribuição percentual das observações por cada estado brasileiro. Esse gráfico destacou que São Paulo concentra a maior parte dos registros, seguido por Minas Gerais e Rio de Janeiro. Essa concentração pode está relacionada a fatores como a densidade populacional, a infraestrutura de distribuição e a presença de maior número de revendas nessas localidades. Além disso, esses estados possuem relevância econômica e forte participação no consumo de combustíveis, o que tende a atrair maior volume de coletas e monitoramentos. Em contrapartida, unidades da federação com menor população e menor integração logística registraram percentuais mais reduzidos, refletindo tanto características demográficas quanto particularidades regionais na comercialização de combustíveis.

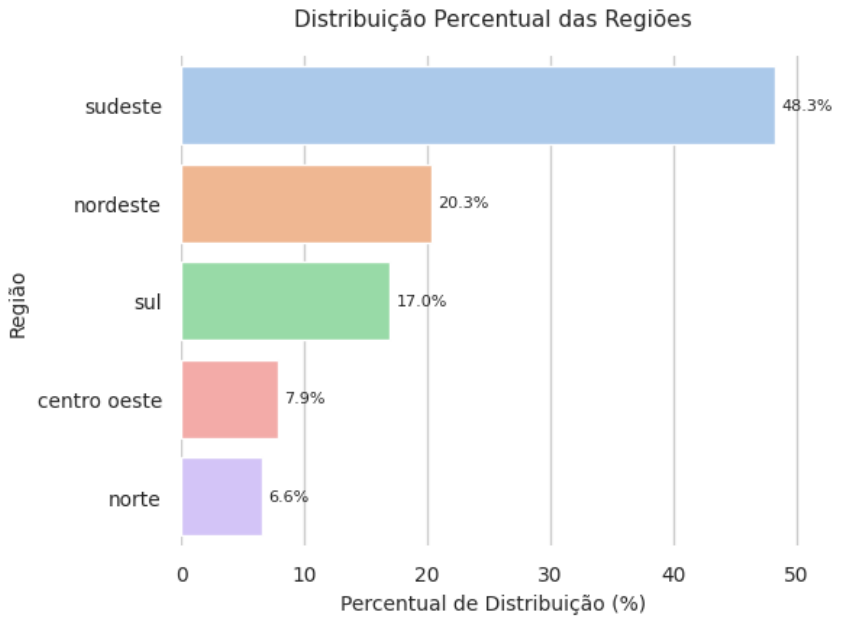

Gráfico 4. Distribuição percentual das coletas por região.
Fonte: A autora (2025).

Já Gráfico 4, mostra a distribuição percentual das coletas por região. Observa-se que o Sudeste concentrou a maior parte dos registros com cerca de 48,3% seguido pelo Nordeste com 20,3% e Sul representada por 17,0%. Por outro lado, as regiões Centro-Oeste e Norte, com 7,9% e 6,6% respectivamente, apresentaram uma baixa representatividade no conjunto de dados, o que pode estar relacionado a aspectos como distância de centros urbanos, menor número de postos ou limitações logísticas.

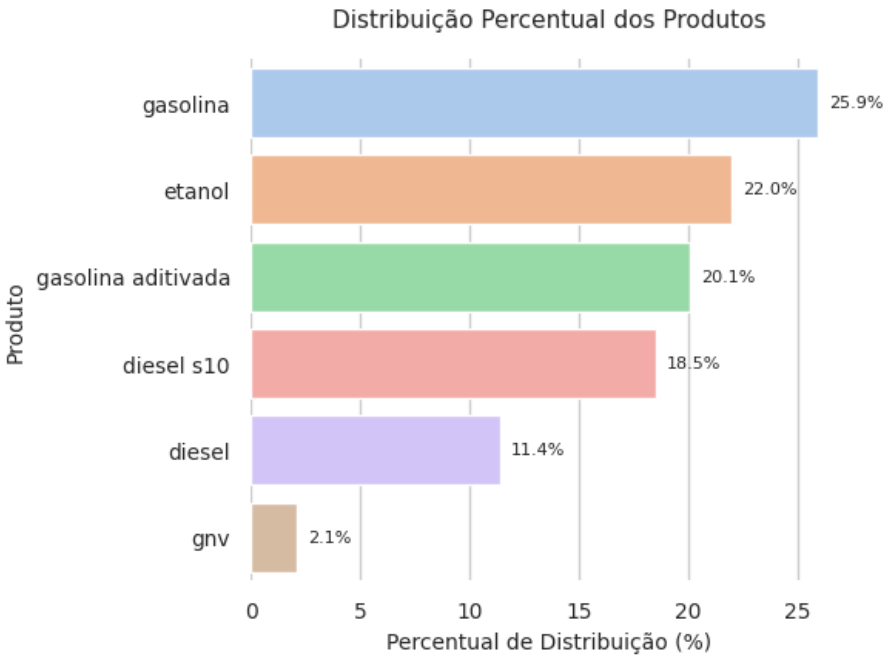

Gráfico 5. Distribuição percentual dos combustíveis.
Fonte: A autora (2025).

O Gráfico 5, apresenta a distribuição percentual dos diferentes combustíveis comercializados no período analisado. A gasolina comum foi o produto mais representativo, correspondendo a 25,9%, seguido pelo etanol, com 22,0%, e a gasolina aditivada, com 20,1%. Por outro lado, o diesel S10 aparece com 18,5%, enquanto o diesel comum representou 11,4% da distribuição. Por último, o Gás Natural Veicular (GNV) teve a menor participação, com apenas 2,1%, o que poderia indicar uma menor demanda do mercado em relação aos outros produtos.

Essa distribuição evidencia a predominância dos combustíveis derivados de petróleo no mercado, principalmente pela a gasolina e pelo diesel. Apesar da presença de alternativas mais sustentáveis como etanol e GNV, esses produtos ainda possuem participação inferior. A maior representatividade dos combustíveis fósseis pode estar relacionada à composição da frota de veículos nacional, à ampla distribuição de abastecimento e aos preços praticados no período analisado.

5.2. Preço Médio dos Combustíveis

Esta etapa contempla a representação gráfica dos valores médios praticados, a partir dos registros consolidados para o período analisado.

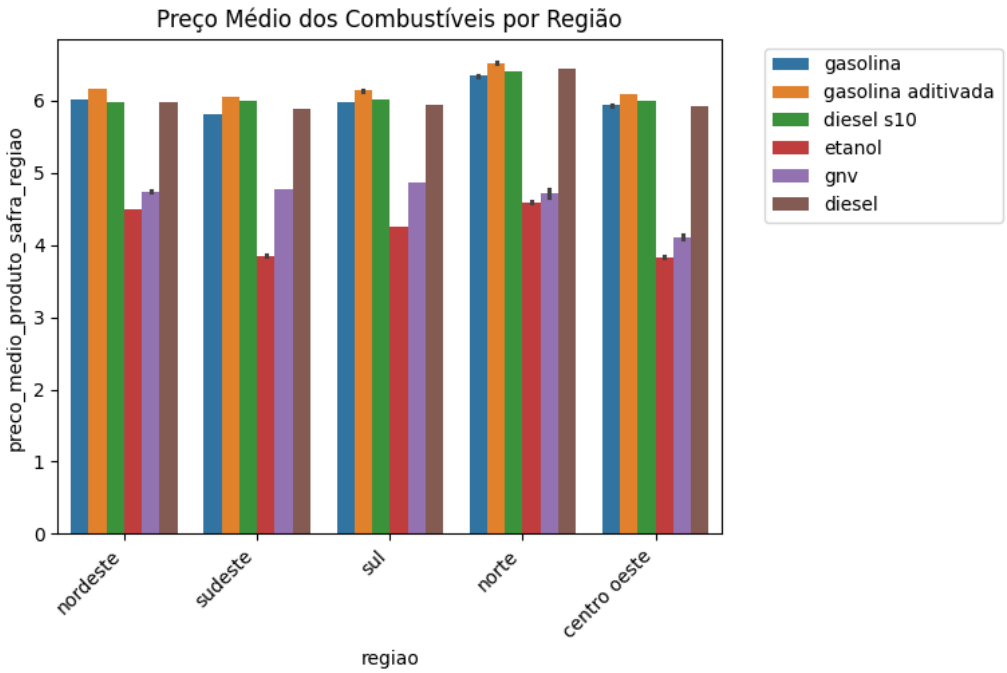

Gráfico 6. Preço médio dos combustíveis por região e tipo de produto.
Fonte: A autora (2025).

O Gráfico 6 acima, compara o preço médio dos principais combustíveis entre as regiões. A análise apresenta que para a maioria dos produtos, a região Sudeste praticou preços médios levemente mais inferiores em relação aos demais, enquanto a região Norte refletiu valores um pouco mais elevados. Essas diferenças podem estar relacionadas a fatores como custo de transporte do centro de produção até a revenda, estrutura tributária estadual e competitividade local entre distribuidoras. Além disso, aspectos como a disponibilidade de infraestrutura de abastecimento, a distância em relação aos polos de

refino e a demanda regional por determinados combustíveis também podem influenciar as variações observadas, reforçando a importância de considerar o contexto logístico e econômico de cada localidade na análise comparativa dos preços.

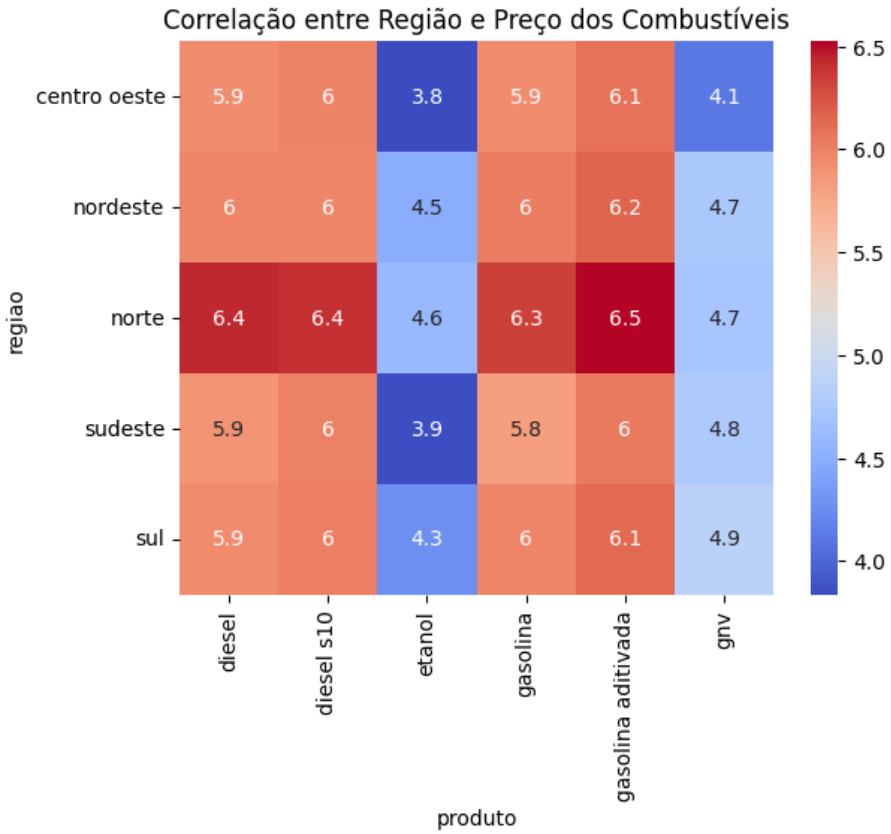

Gráfico 7. Média do valor de venda por produto e região.
Fonte: A autora (2025).

Complementando esse contexto, o gráfico 7, apresenta a média dos valores de venda por produto em cada região do Brasil por meio do gráfico de calor. O Norte concentra os maiores preços médios para todos os combustíveis, principalmente para a gasolina aditivada R$ 6,50 e o diesel S10 R$ 6,40

Por outro lado, os menores valores médios de venda foram observados nas regiões do Centro-Oeste e Sudeste, mais especificamente para o etanol, que atingiu R$ 3,80 e R$ 3,90, respectivamente. O GNV apresentou preços mais baixos de venda em todas as regiões, em relação a maioria dos preços dos outros produtos, sendo mais barato no Centro-Oeste R$ 4,10 e mais caro no Sul R$ 4,90.

Em geral, os dados demonstram que os preços médios de combustíveis no Brasil não são uniformes, variando em cada região, reforçando a importância da análise geográfica para entender o comportamento do mercado.

5.3. Análise Relativa dos Preços por Região

Será apresentada, a seguir, uma análise da relação entre o preço relativo e o valor de venda dos combustíveis, com a visualização dos dados por região em cada gráfico.

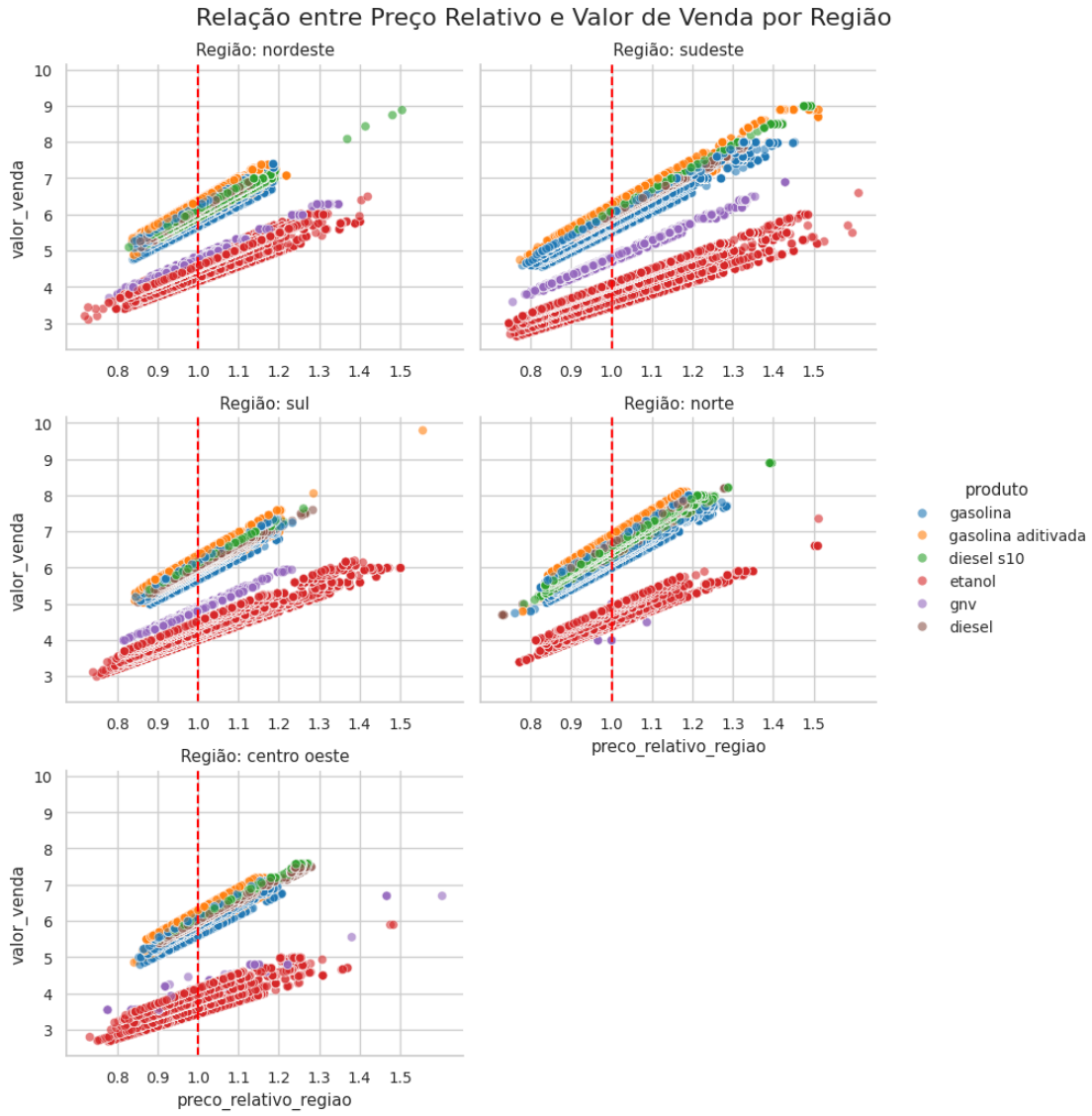

Gráfico 8. Distribuição percentual das coletas por região
Fonte: A autora (2025).

O Gráfico 8, representa esta relação como a razão entre o preço praticado e a média nacional e o valor de venda dos combustíveis. A linha vermelha tracejada em $x = 1$ serve como referência para identificar variações regionais nos preços, pois ela representa a média nacional.

Nas regiões Nordeste, Norte e Centro-Oeste, os registros se concentram entre $x = 0,90$ e $x = 1,20$, o que poderia indicar um maior alinhamento com a média nacional. Nessas regiões, predominam produtos como etanol e diesel comum, e os seus preços de venda ficam, em geral, entre R$ 4,00 e R$ 5,50. A maior presença desses combustíveis, com valores mais baixos por litro, contribui para manter os preços relativos próximos de 1.

No Sudeste, a maior parte dos pontos aparece do lado direito da linha, principalmente entre $x = 1{,}00$ e $x = 1,30$, sugerindo que os preços praticados nessa região, comparando com as demais regiões, costumam ser os mais altos do país. Em relação aos

valores de venda, estão mais concentrados no intervalo de R$ 4,00 e R$ 7,50. Os produtos que mais influenciam são a gasolina, gasolina aditivada e o disel S10 que possuem valores mais elevados entre os combustíveis.

A região Sul possui um comportamento em seus preços acima da média nacional, com uma agregação dos pontos mais à direita da linha, especialmente entre $x = 1{,}00$ e $x = 1,20$, indicando que os preços praticados neste local geralmente fica acima da média nacional. No gráfico, observa-se que os registros de gasolina e diesel S10, combustíveis com valor de venda entre R$ 5,00 e R$ 6,50.

Essa análise sugere que a composição dos combustíveis e as diferenças regionais de infraestrutura e logística podem influenciar os preços praticados em cada localidade.

5.4. Evolução dos Preços ao Longo do Tempo

O estudo abaixo, apresenta a variação do valor de venda médio dos combustíveis ao longo dos meses de 2024.

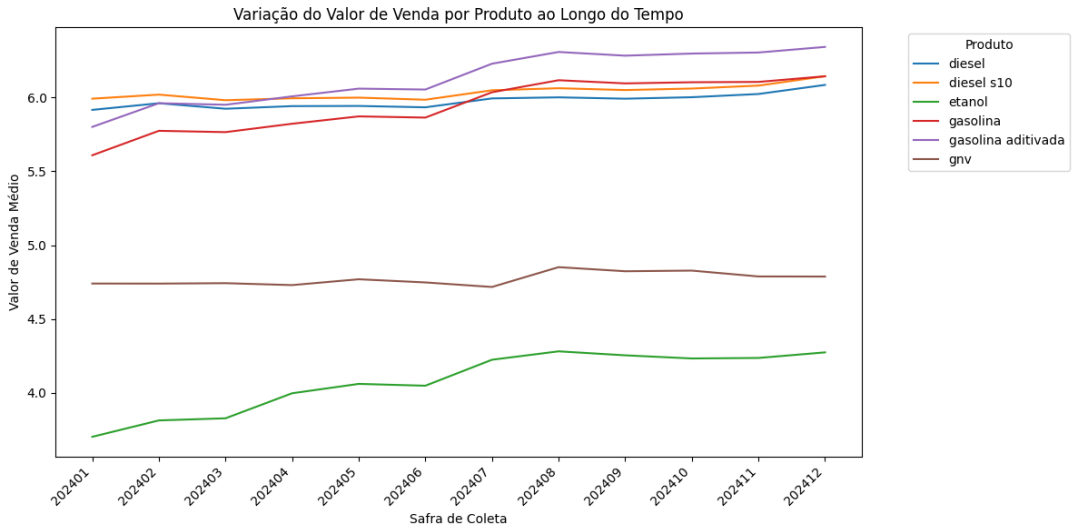

Gráfico 9. Variação do valor de venda médio por produto ao longo de 2024.
Fonte: A autora (2025).

No Gráfico 9, acima, a gasolina aditivada manteve os maiores valores médios de venda durante todo o período analisado, seguida pela gasolina comum, diesel S10 e diesel comum, que apresentaram valores próximos entre si. O produto com menor valor médio entre os combustíveis líquidos foi o etanol, embora tenha apresentado crescimento gradual ao longo do tempo. Já o GNV apresentou a menor variação e o menor valor médio ao longo do ano.

De forma geral, todos os produtos registraram uma tendência de aumento nos preços ao longo do ano, com um aumento entre os meses de julho e agosto, o que possivelmente reflete ajustes sazonais ou variações no custo de produção e distribuição. Além disso, as diferenças observadas entre os produtos, tanto em níveis médios quanto em variações ao longo do tempo, evidenciam comportamentos distintos entre os produtos no conjunto de dados analisado, reforçando a importância de uma abordagem baseada em múltiplas variáveis para captar as dinâmicas do mercado.

5.5. Análise de Outliers e Alertas com ydata-profiling

Para complementar a análise exploratória e avaliar possíveis distorções nas variáveis numéricas, foi realizado um estudo de *outliers*, com o objetivo de identificar valores extremos.

A avaliação foi executada através da ferramenta *ydata*, que aplica critérios estatísticos clássicos, como intervalo interquartil e medidas de dispersão, para identificar os potenciais registros discrepantes. No entanto, nenhuma das variáveis da base foram sinalizadas com alertas de valores extremos ou distribuições altamente assimétricas.

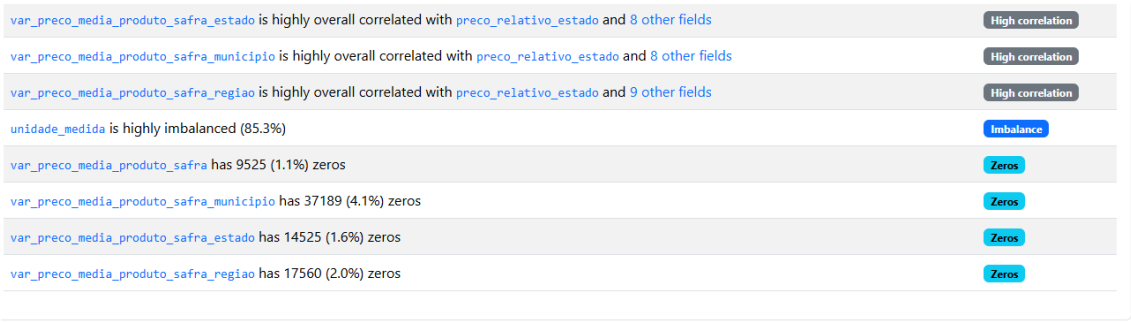

Figura 2. Captura dos alertas indicado na ferramenta ydata

Fonte: A autora (2025).

Esse resultado indica a consistência da base e aponta que os dados disponíveis refletem um padrão relativamente estável, sem distorções que necessitassem de tratamentos adicionais antes da aplicação do algoritmo de clusterização. Dessa forma, optou-se por manter todos os registros na etapa seguinte do projeto.

Além da análise dos *outliers*, o relatório gerado pelo *ydata*, apontou 40 alertas que serve para identificar potenciais problemas estatísticos na base. Os três tipos de alertas apresentado foram: *high correlation*, *imbalance* e *zeros* como pode ser visualizado na (Figura 2).

- **High correlation**: variáveis de variação de preço médio foram calculadas em diferentes granularidades (estado, município e região) apareceram altamente correlacionadas com a coluna de preço relativo. Esse resultado era previsto, pois as métricas derivam das mesmas séries de preços. Para evitar redundância e reduzir o risco de distorção das distâncias no *K-Means*, foram mantidas apenas a granularidade por região, alinhada ao objetivo da clusterização.
- **Imbalance**: a variável unidade_medida indicou um alto desbalanceamento, cerca de 85,3% dos registros em litros, o que também era esperado, visto que apenas o GNV é reportado em metro cúbico e tem uma menor volumetria na base de dados. Essa variável não foi utilizada no modelo de clusterização e não foi necessário aplicar um tratamento adicional.
- **Zeros**: variáveis de variação de preço var_preco_media_produto_*, apresentaram pequenas proporções de zeros, entre 1,1% e 4,1%, indicando períodos de estabilidade no preço médio, o que é esperado, porque essas colunas fornecem informações sobre estabilidade dos preços.

6. Resultados

Esta seção apresenta os resultados obtidos após a aplicação do algoritmo de clusterização *K-means* na base de dados construída, conforme descrito nas etapas anteriores. As análises foram organizadas em duas abordagens complementares: a segmentação por região e a segmentação por produto. O objetivo é identificar agrupamentos com características semelhantes, permitindo explorar padrões estruturais no comportamento dos preços dos combustíveis. Por fim, são discutidas as características de cada grupo identificado, com base nas variáveis utilizadas no modelo.

6.1. Análise por Região

Com base nas evidências identificadas na análise exploratória, foi aplicado o algoritmo de clusterização *K-Means* para identificar grupos homogêneos das regiões com base em variáveis numéricas e a categórica. A definição do número de agrupamentos foi realizada utilizando o método do cotovelo (*Elbow Method*), que indicou a existência de três *clusters* como ideal.

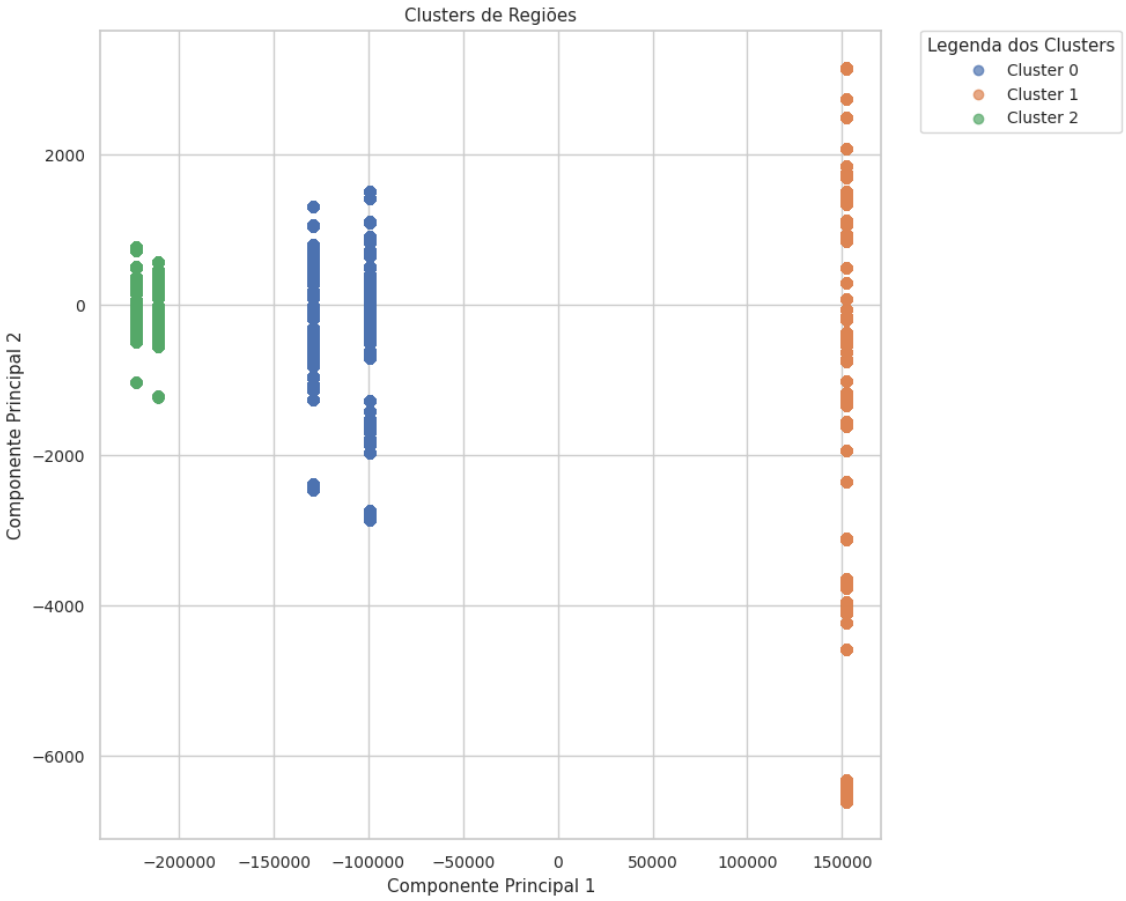

Gráfico 10. Distribuição dos clusters de região após PCA.
Fonte: A autora (2025).

As variáveis numéricas selecionadas foram: preço médio por produto/safra, variação do preço médio por safra, preço relativo, quantidade de coletas, quantidade de revendas e número de produtos por revenda, todas na granularidade por região. A escolha

dessas variáveis considerou suas características relacionadas ao comportamento regional dos combustíveis, como o nível de preços praticados, sua estabilidade ao longo do tempo, o posicionamento relativo no mercado e a estrutura da rede de revendas em cada região. As demais variáveis da base, como aquelas associadas as características estaduais, municipais ou informações específicas por revenda, foram desconsideradas por apresentarem maior granularidade ou estarem mais conectadas as variações locais, o que poderia introduzir ruídos e dificultar a identificação de padrões entre as regiões. A variável categórica *região* foi transformada em númerica por meio de *One-Hot Encoding*.

As variáveis numéricas foram padronizadas com *StandardScaler*, e os resultados da clusterização foram visualizados por meio de uma redução de dimensionalidade via *PCA*. O Gráfico 10, apresenta os agrupamentos obtidos, onde é possível observar separações bem definidas entre os grupos.

A qualidade da segmentação foi avaliada por meio do *Silhouette Score*, com valor médio de aproximadamente 0,929. Para esse cálculo, foi utilizada uma amostra aleatória de 50.000 registros. Essa decisão foi motivada pela necessidade de otimizar o desempenho computacional no ambiente *Google Colab*, pois os testes realizados com a base completa (com mais de 890 mil registros) resultaram em alto custo de processamento, chegando a interromper a execução do ambiente. De acordo com Pavlopoulos et al. (2024), em cenários que envolvem grandes volumes de dados, a utilização de uma amostra é uma prática recomendada para estimar o *Silhouette Score* de forma eficiente. Neste estudo, a amostra selecionada representou uma parcela significativa da base e foi suficiente para fornecer uma estimativa confiável da coesão interna e da separabilidade dos grupos, sem comprometer a confiabilidade da métrica.

A seguir, as características principais de cada *cluster* identificado são explicadas:

- *Cluster* 0 (cor azul): representado pelas regiões Sul e Nordeste. Esse grupo se destaca pela alta variabilidade nos preços médios por produto e safra, o que poderia estar relacionado a um número mais reduzido de coletas e menor presença de revendas nesses locais. Cerca de 37,26% dos registros está presente neste grupo.
- *Cluster* 1 (cor laranja): formado apenas por registros do Sudeste e compõe 48,29% dos dados. Apresenta maior homogeneidade, com alto volume de coletas e revendas (acima de 74%) e preços mais estáveis que são relacionados a uma estrutura de mercado mais consolidada. Sua separação é bem definida na projeção 2D, na imagem anterior, reforçando a robustez desse agrupamento.
- *Cluster* 2 (cor verde): inclui as regiões Norte e Centro-Oeste, correspondendo a 14,45% dos registros. Apesar do menor volume de dados, esse grupo se diferencia por apresentar preços acima da média nacional e perfil de revenda que pode apresentar características específicas em relação às demais regiões.

De forma geral, os resultados apontam disparidades estruturais importantes entre as regiões brasileiras em relação a comercialização de combustíveis. A combinação da análise exploratória com métodos de clusterização proporcionou uma compreensão mais profunda dos padrões regionais, contribuindo para a formulação de diagnósticos e estratégias específicas para cada contexto geográfico.

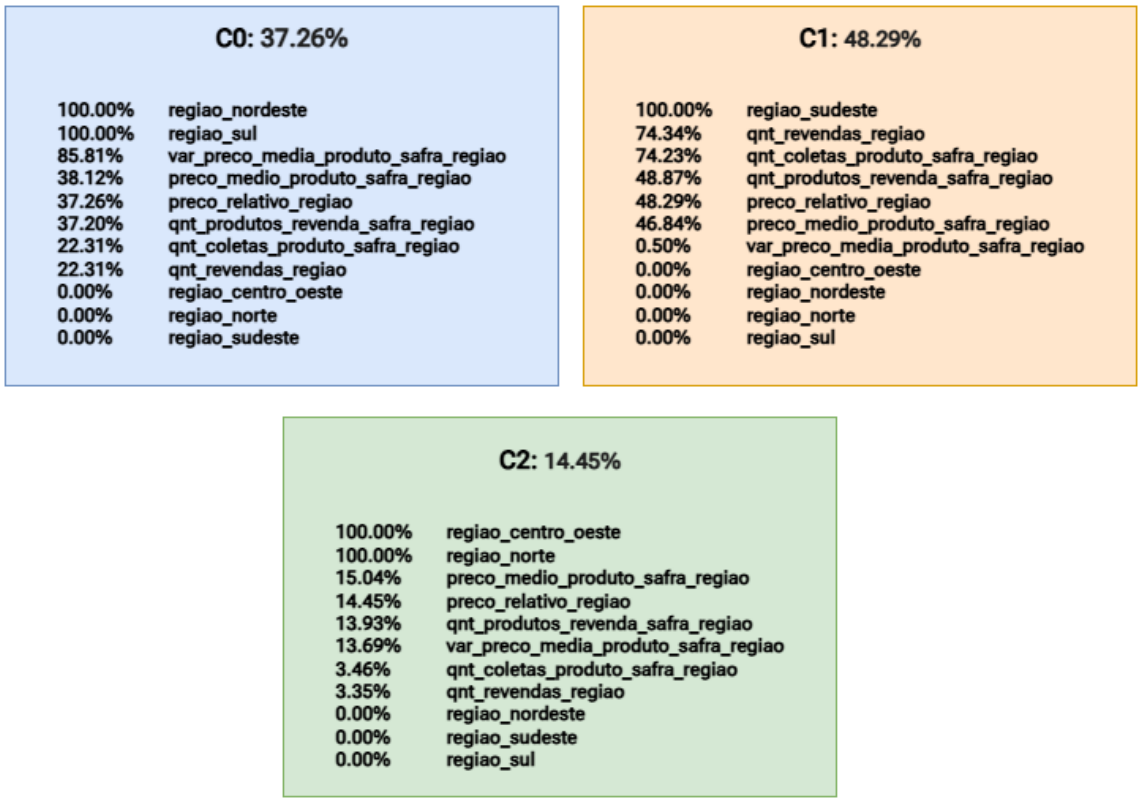

Figura 3. Detalhamento dos clusters
Fonte: A autora (2025).

A imagem 3, possui diversas porcentagens que foram obtidas a partir de dois tipos de cálculo. O primeiro corresponde ao percentual de registros de cada grupo em relação ao total da amostra, ou seja, foi calculada a proporção das informações pertencentes a cada *cluster*, resultando em C0 com 37,26% dos registros, C1 com 48,29% e C2 com 14,45%. O segundo cálculo foi realizado para compreender a distribuição das variáveis numéricas dentro de cada *cluster*. Para isso, foi considerada a soma dos valores de cada variável em cada agrupamento, dividida pela soma total dessa mesma variável na base completa.

Essa duas abordagens possibilitam identificar o tamanho relativo dos grupos, além da representatividade de cada *cluster* em relação às variáveis analisadas, oferecendo *insights* mais profundos sobre as características predominantes em cada segmento formado.

Os percentuais para cada *cluster*, indicam que o C1 é o grupo dominante, concentrando quase metade do conjunto de dados com características semelhantes. O C0 apresenta uma parcela expressiva, embora menor que o C1, enquanto o C2 representa o menor agrupamento, reunindo um subconjunto específico de registros.

Analisando os percentuais da variável preco_medio_produto_safra_regiao, observa-se que os valores ficaram próximos à distribuição dos registros totais: C0 com 38,12%, C1 com 46,84% e C2 com 15,04%. A variável preco_relativo_regiao manteve exatamente os mesmos percentuais da volumetria total, o que indica alinhamento entre o comportamento do preço relativo e a distribuição geral dos dados.

Por outro lado, a variável var_preco_media_produto_safra_regiao apresentou uma concentração desbalanceada: 85,81% da soma total dessa variável se concentrou no *cluster* C0, enquanto apenas 0,50% apareceu no C1 e 13,69% no C2. Essa disparidade ocorre porque essa variável representa a variação entre o valor_venda e o preco_medio_produto_safra_regiao. Como se trata de uma diferença, pode resultar em valores negativos ou positivos e ao se considerar a soma dentro de cada *cluster*, essa oscilação impacta o percentual acumulado. Assim, o C0 concentrou a maior parte das variações observadas.

A variável qnt_coletas_produto_safra_regiao reforça o domínio do *cluster* C1, que representa 74,23% da soma total, enquanto C0 ficou com 22,31% e C2 com apenas 3,46%. Essa distribuição semelhante é vista em qnt_revendas_regiao, com o C1 somando 74,34%, o C0 com 22,31% e o C2 com 3,35%. Já a variável qnt_produtos_revenda_safra_regiao refletiu uma divisão mais equilibrada: C0 com 37,20%, C1 com 48,87% e C2 com 13,93%.

Em relação à variável categórica regiao, após a aplicação do *One Hot Encoding*, foi observado que os *clusters* foram formados de maneira segmentada: o C0 possui exclusivamente as regiões Nordeste e Sul (ambas com 100%), o C1 representa apenas a região Sudeste (100%) e o C2 concentra os registros das regiões Norte e Centro-Oeste (100% em cada).

Essa distribuição reforça a interpretação de que o agrupamento por região gerou *clusters* altamente coerentes com características específicas tanto em termos de preço quanto de volumetria. O C1 representa uma região densa em volume e estrutura comercial, Sudeste, o C0 agrega regiões de comportamento semelhante mas com menor variabilidade de volume, Sul e Nordeste, enquanto o C2, mesmo sendo o menor, destaca-se por apresentar padrões próprios de preço e cobertura nas regiões Norte e Centro-Oeste.

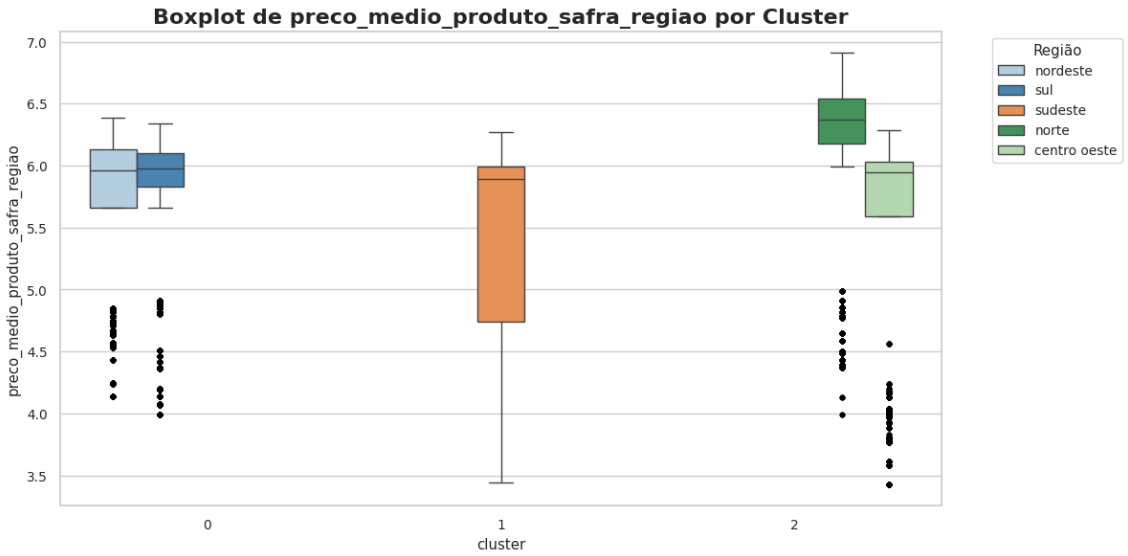

Gráfico 11. Distribuição dos preços médios por produto e safra em cada cluster.
Fonte: A autora (2025).

Os gráficos 11, 12 e 13 apresentam a distribuição estatística por meio do gráfico boxplot para as variáveis preco_medio_produto_safra_regiao, preco_relativo_regiao

e qnt_coletas_produto_safra_regiao, respectivamente, em cada um dos *clusters* gerados para a visão regional. A escolha dessas variáveis se justifica por refletirem três dimensões complementares da análise: o preço médio praticado ao longo das safras, o comportamento relativo dos preços em comparação à média nacional e o volume de coletas realizadas em cada região. Essa combinação permite observar variações de tendência, dispersão e representatividade dos dados dentro de cada *cluster*.

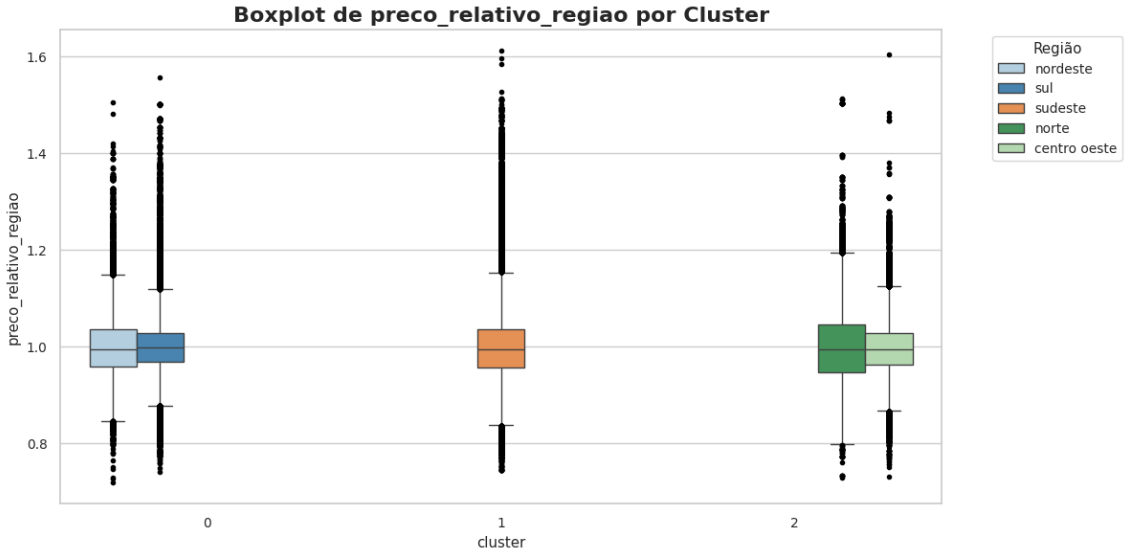

Gráfico 12. Distribuição dos preços relativos por região em cada cluster.
Fonte: A autora (2025).

No gráfico 11, observa-se que o *cluster* C0 (representado pelas regiões Nordeste e Sul) apresenta caixas com tamanhos semelhantes e concentradas em torno de valores entre 5,8 e 6,2. Isso sugere uma menor variabilidade nos preços médios dessas regiões, indicando certa estabilidade. Já no *cluster* C1, que corresponde a região Sudeste, indica uma maior amplitude entre os *boxplots*, com valores que vão aproximadamente de 3,5 até 6,3. A mediana se encontra próxima a 5,9, e a caixa mais comprida entre os *clusters* reforça a diversidade de comportamentos de preço dentro dessa região. Por fim, o *cluster* C2, que agrega as regiões Norte e Centro-Oeste, apresenta valores mais elevados e concentrados entre 6,0 e 6,8, com a mediana mais alta entre os três *clusters*. Isso poderia indicar que essas regiões praticaram preços médios mais altos no acumulado de 2024.

O Gráfico 12, apresenta os valores do preco_relativo_regiao, ou seja, a razão entre o preço praticado em determinada região e a média nacional para o mesmo produto. Aqui, todos os *clusters* apresentam mediana próxima a 1, o que indica que os preços praticados em geral não se distanciam tanto da média nacional. Entretanto, o *cluster* C2 apresenta uma maior desproporção, com regiões que possuíram seus preços acima da média. O C0 mantém uma distribuição semelhante à do C1, porém com caudas levemente mais alongadas para valores abaixo da média.

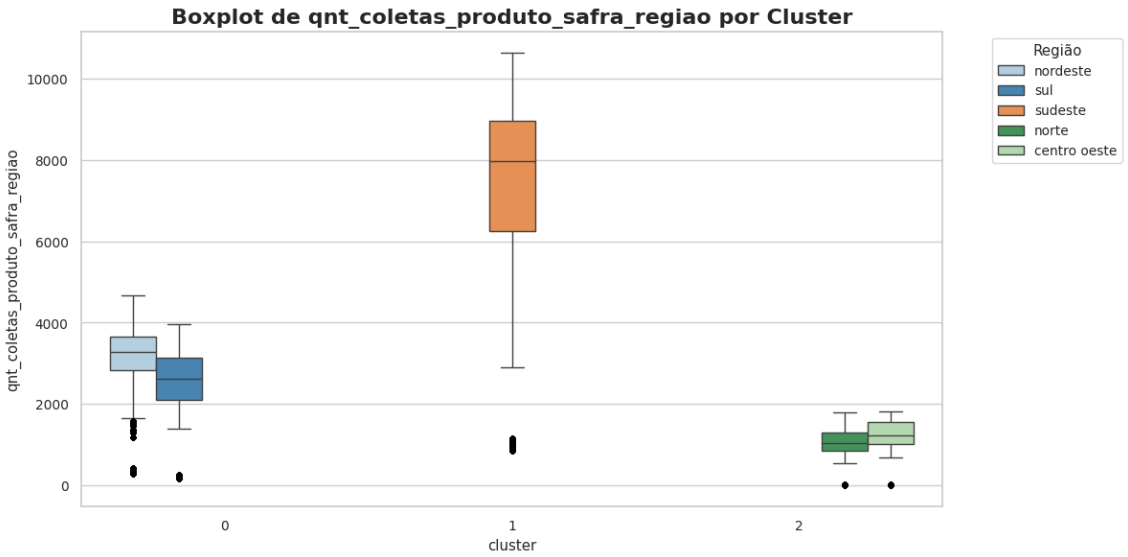

Gráfico 13. Distribuição da quantidade de coleta por região em cada cluster.
Fonte: A autora (2025).

Por fim, o gráfico 13 mostra a variável qnt_coletas_produto_safra_regiao, que representa o volume de amostras coletadas por região. O *cluster* C1, contém dos dados da região Sudeste e possui uma distribuição mais elevada, com mediana próxima a 8000 coletas e valores que ultrapassam os 10 mil.

Os *clusters* C0 e C2 apresentam menor volume de coletas, com mediana entre 1000 e 3500, sendo o C2 o menor entre as variações. Esses resultados reforçam a representatividade do Sudeste nos dados e ajudam a justificar a dominância estatística do *cluster* C1 observada nas analises anteriores.

De forma geral, os gráficos boxplots reforçam os padrões identificados na clusterização: o C1 com maior volume e maior variabilidade interna, o C0 com preços moderados e consistentes e o C2 com preços mais elevados e menor representatividade em relação ao volume das coletas.

6.2. Análise por Produto

Assim como a visão de Região, para Produtos, também foi aplicado o algoritmo de clusterização *K-Means* para visualizar os grupos semelhantes a partir dos parâmetros utilizados. Para isso, também foi aplicado o método do cotovelo (*Elbow Method*) e apontou como o ideal, a existência de 04 *clusters*.

As variáveis numéricas utilizadas incluíram: preço médio, variação do preço médio, preço relativo, quantidade de coletas, quantidade de produtos por revenda, sendo todas segmentadas por produto. A seleção dessas variáveis foi baseada na relevância para representar características fundamentais em cada combustível, buscando coletar aspectos como nível de preço, estabilidade, posicionamento relativo no mercado e volume de registros. Esses fatores foram considerados mais apropriados ao contexto da análise por produto, uma vez que refletem atributos referentes ao comportamento individual de cada combustível. Variáveis segmentadas por região, estado ou município foram desconsideradas neste agrupamento, pois poderiam introduzir distorções associadas a fatores

geográficos e comprometer a identificação de padrões relacionados ao comportamento dos combustíveis.

A variável categórica *produto* foi transformada por meio de *One-Hot Encoding*. Já as variáveis numéricas foram padronizadas com *StandardScaler* e os resultados da clusterização foram visualizados por meio de uma redução de dimensionalidade via *PCA*.

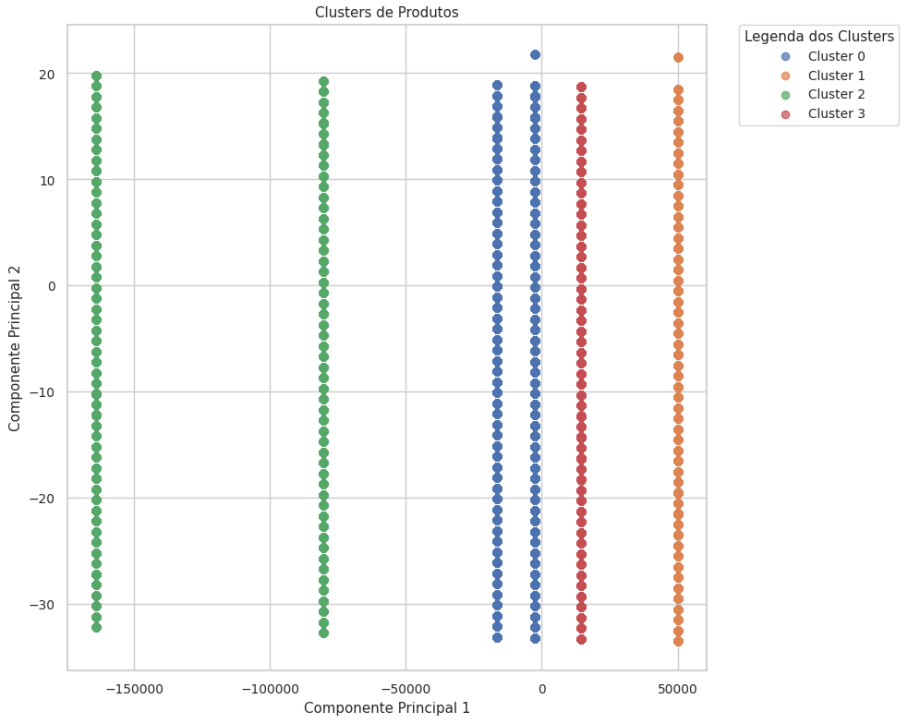

Gráfico 14. Distribuição dos clusters após PCA.
Fonte: A autora (2025).

No Gráfico 14, apresenta a segmentação dos produtos de combustíveis por meio do algoritmo *K-Means* e essa clusterização resultou em quatro grupos distintos, cada um representando um perfil específico de produto com base em suas características de comercialização. A seguir, as características principais de cada *clusters* identificado são explicadas:

- *Cluster* 0 (cor azul): possui em sua maior parte o diesel S10 e a gasolina aditivada com 47,9% e 52,0%, respectivamente, indicando que esses dois produtos compartilham atributos semelhantes no conjunto analisado.
- *Cluster* 1 (cor laranja): é composto apenas pela gasolina comum, indicando que o seu comportamento é isolado em relação aos demais produtos.
- *Cluster* 2 (cor verde): representa principalmente o Gás Natural Veicular (GNV), com 15,6% de ocorrência, enquanto os demais produtos não estão presentes nesse grupo, o que pode indicar um padrão muito específico associado ao GNV.
- *Cluster* 3 (cor vermelha): concentra apenas o etanol, o que poderia reforçar suas particularidades em termos de perfil de comercialização.

Essa segmentação permite identificar como os produtos se agrupam segundo suas características operacionais e de mercado, oferecendo *insights* para análises mais direci-

onadas sobre precificação, necessidade e políticas de incentivo conforme o tipo de combustível.

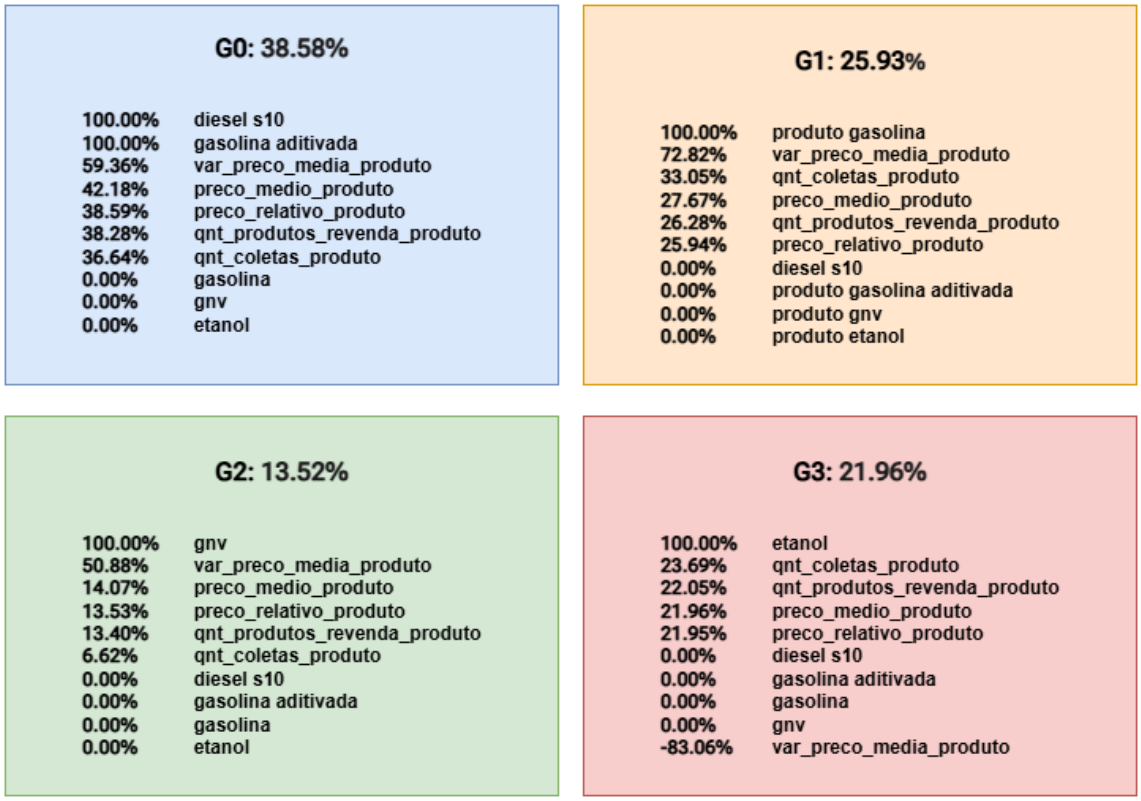

Figura 4. Detalhamento dos clusters
Fonte: A autora (2025).

Conforme metodologia já descrita, a avaliação da qualidade dos agrupamentos também foi realizada por meio do *Silhouette Score*, obtendo valor médio de 0,844 na segmentação por produto. Esse resultado indica uma boa separação entre os grupos formados, com coesão interna satisfatória, embora um pouco inferior ao observado na análise por região.

A análise dos *clusters* gerados a partir da visão por produto indicou agrupamentos bem definidos, nos quais cada grupo concentrou na sua maioria um único tipo de combustível. Essa segmentação possibilitou observar padrões distintos de comportamento em relação à variação de preços, quantidade de coletas e presença por revenda.

A Figura 4, apresenta a distribuição percentual da soma de cada variável por *cluster*, para a análise com foco na visão por produto. As porcentagens foram calculadas com base nos mesmos critérios aplicados na análise por região, seção anterior, considerando a proporção de registros e a participação percentual na soma de cada variável numérica por agrupamento.

Os valores representam a frequência relativa de cada variável dentro de cada agrupamento. Observa-se que o *cluster* G0 concentrou 38,58% dos registros, seguido por 25,93% no G1, 13,52% no G2 e por fim, 21,96% no G3 . Esse cenário revela que os da-

dos foram relativamente bem distribuídos entre os *clusters*, com o G0 aparecendo como o maior grupo, enquanto o G2 representou como o menor.

A variável preco_medio_produto demonstrou uma distribuição mais alinhada à volumetria geral, com destaque para o G0 (2,18% e o G1 27,67%, enquanto o G2 (14,07%) e o G3 21,96% ficaram com a concentração mais abaixo. Similar a este comportamento, também ocorre com a variável preco_relativo_produto, onde os valores acompanharam a lógica da volumetria: G0 com 38,59%, G1 com 25,94%, G2 com 13,53% e G3 com 21,95%.

Por outro lado, a variável var_preco_media_produto refletiu um comportamento assimétrico entre os *clusters*. O G1 concentrou 72,82% do total da variável, indicando que a maior parte das variações foi observada nesse grupo. Já o G0 ficou com 59,36%, o G2 com 50,88% e o G3 apresentou um valor negativo de -83,06%, o que indica que a influência de oscilações negativas nos cálculos, dado que a variável representa a diferença entre o valor de venda e a média por produto.

Em relação à variável qnt_coletas_produto, o *cluster* G0 manteve a maior volumetria de registros com 36,64%, seguido por G1 33,05%, G3 23,69% e G2 (6,62%). Já a variável qnt_produtos_revenda_produto também apresentou um padrão similar: G0 com 38,28%, G1 com 26,28%, G3 com 22,05% e G2 com 13,40%.

As variáveis categóricas de produto indicou que cada *cluster* concentrou tipos específicos de combustível. O *cluster* G0 representa apenas os produtos de diesel S10 e gasolina aditivada ambos com 100%, sugerindo uma forte segmentação por tipo de combustível. O G1 também apresentou a gasolina com 100%. Já o G2 é relacionado ao gnv, enquanto o G3 concentra 100% de registros do etanol.

Essas segmentações indicam que o agrupamento dos dados pela granularidade de produto gerou *clusters* com forte coerência em relação ao tipo de combustível, com cada grupo concentrando predominantemente apenas um ou dois produtos, o que facilita a análise de comportamento específico por segmento do mercado de combustíveis.

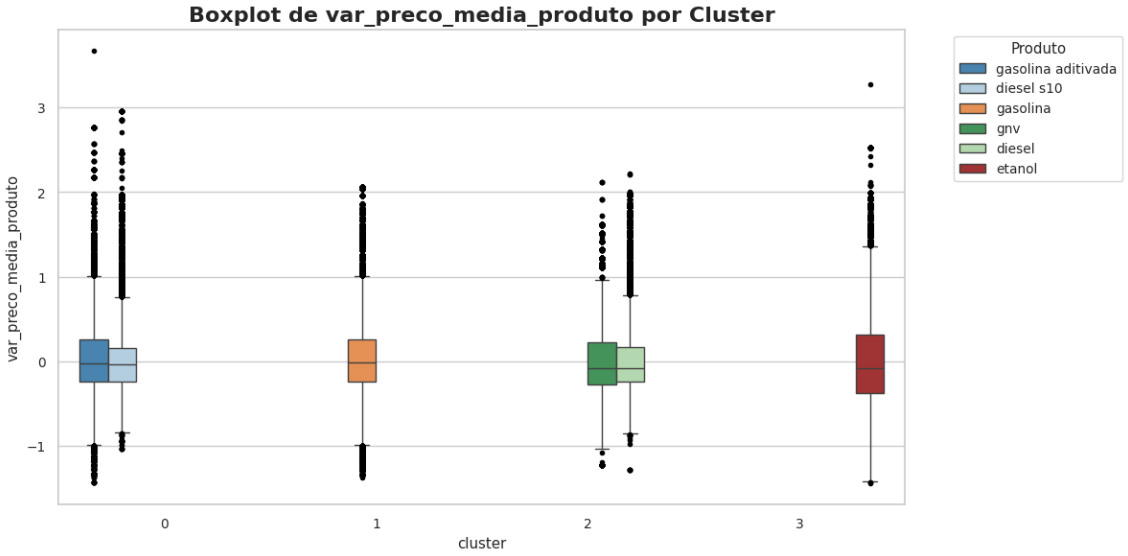

Gráfico 15. Distribuição da variação do preço médio por produto em cada cluster.
Fonte: A autora (2025).

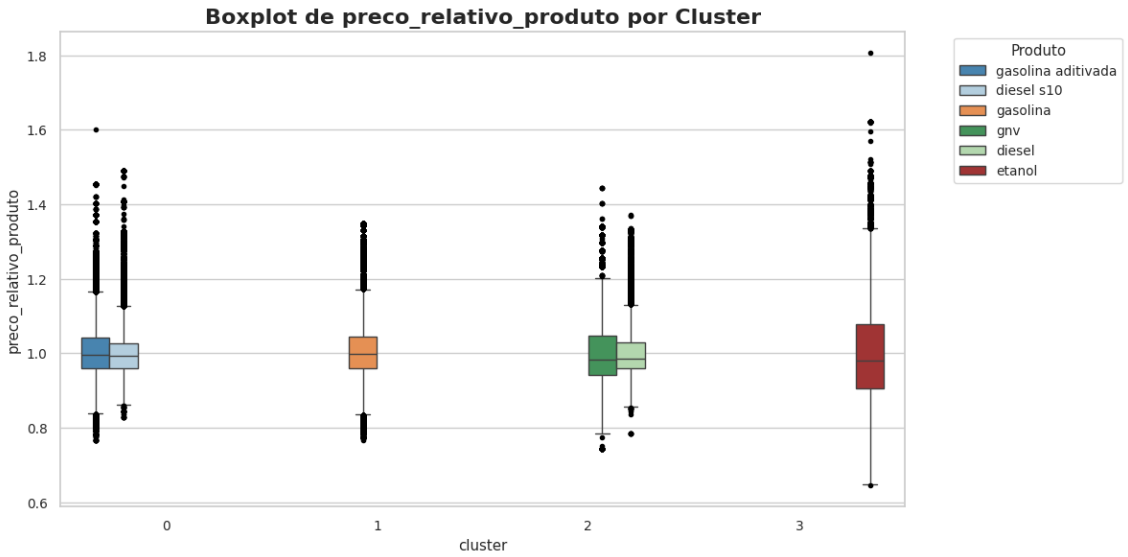

Gráfico 16. Distribuição do preço relativo por produto em cada cluster.
Fonte: A autora (2025).

A distribuição das variáveis var_preco_media_produto e preco_relativo_produto por *cluster*, são apresentadas nos gráficos 15 e 16, com diferenciação dos combustíveis por cor. A escolha dessas variáveis busca representar dois fatores relevantes na comparação entre os produtos: a intensidade das variações nos preços médios, expressa pela dispersão dos valores dentro de cada grupo, e a posição relativa dos produtos em relação à média geral, permitindo observar quais tendem a ser mais caros ou mais baratos no mercado. Essas duas abordagens possibilitam a identificação de padrões dentro de cada *cluster*.

O *cluster* G3, é composto pelos dados do etanol e se destaca por apresentar a maior amplitude de valores em ambas as variáveis, indicando a elevada volatilidade deste produto. Por outro lado, os demais grupos mantêm distribuições mais concentradas e semelhantes em ambos os gráficos. O *cluster* G0, composto por gasolina aditivada e diesel S10, e o *cluster* G1 composto pela gasolina, demonstram comportamento mais estável, com menor variação e a distribuição dos valores estão mais equilibradas em relação a mediana. Já o *cluster* G2, formado por GNV e diesel comum, apresenta uma leve dispersão, porém dentro de um intervalo mais estreito, sem grandes desvios.

Esses resultados reforçam a robustez da segmentação obtida pela clusterização, que não apenas separa os registros por tipo de produto, mas também separa os grupos com diferentes níveis de oscilação de preços. A coerência entre os padrões de variação e de valores relativos sugere que os *clusters* coletaram características estruturais relevantes do mercado de combustíveis, podendo auxiliar políticas especificas por produto.

7. Conclusões

A aplicação da técnica de clusterização *K-means* permitiu identificar padrões distintos de comportamento tanto entre as regiões brasileiras quanto entre os tipos de combustíveis comercializados em 2024. Na análise regional, observou-se a formação de *clusters* coesos, com separações bem definidas entre Sudeste, Sul/Nordeste e Norte/Centro-Oeste, refletindo desigualdades estruturais e diferenças logísticas que impactam diretamente nos preços e na presença de revendas.

Na análise por tipo de produto, os agrupamentos revelaram uma segmentação natural entre os diferentes tipos de combustíveis, com destaque para o comportamento isolado do etanol e do GNV, além da aproximação entre a gasolina aditivada e o diesel S10. As métricas de avaliação, como o *Silhouette Score*, indicaram boa qualidade em ambos os agrupamentos formados, reforçando a consistência das segmentações obtidas.

Com base nos resultados obtidos, conclui-se que a clusterização é uma abordagem eficiente para revelar padrões latentes em bases extensas e heterogêneas como a da ANP. Os *insights* extraídos neste trabalho contribuem para a compreensão das diferenças estruturais entre regiões e produtos, podendo servir de base para estudos futuros que explorem estratégias comerciais, otimização logística ou investigações sobre comportamento de mercado no setor de combustíveis.

Entre as limitações do artigo, destaca-se a utilização de dados referentes a um único ano (2024), o que restringe a análise de tendências temporais. A escolha por utilizar apenas esse recorte temporal foi motivada por questões práticas de execução, como a limitação de recursos computacionais disponíveis e a necessidade de manter a viabilidade do projeto dentro de ambientes gratuitos, como o *Google Colaboratory*, sem necessidade de infraestrutura computacional robusta ou custos adicionais com serviços em nuvem. Além disso, a decisão de restringir variáveis categóricas visou facilitar a interpretação, mas reduziu o detalhamento mais granular dos *clusters*. Para estudos futuros, uma possibilidade é explorar dados históricos da ANP para análise temporal da dinâmica de preços. Outra alternativa seria a aplicação de algoritmos de agrupamento alternativos, como *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) ou *Gaussian Mixture Models* (GMM), permitindo comparações em relação à abordagem adotada neste trabalho.

Referências

[Alsaifi and outros 2023] Alsaifi, K. and outros (2023). The impact of the russia–ukraine war on global energy markets. *Humanities and Social Sciences Communications*, 10(1):1–12.

[de Andrade and de Souza Campos Rodrigues 2024] de Andrade, V. C. S. and de Souza Campos Rodrigues, G. S. (2024). Análise da formulação da política nacional de biocombustíveis - renovabio: o territorial, o político e o econômico. *Sociedade & Natureza*, 36:e71461.

[Delgado and Gauto 2021] Delgado, F. and Gauto, M. (2021). Composição dos preços de paridade dos combustíveis no brasil. *Revista Conjuntura Econômica*, 75(6):44–48.

[dos Santos Barcellos 2017] dos Santos Barcellos, G. (2017). Visualização de dados obtidos da série histórica de preços de combustíveis no brasil. Monografia de bacharelado, Universidade Federal do Rio Grande do Sul, Porto Alegre.

[Kim 2023] Kim, D. S. (2023). Silhouette analysis for k-means clustering on time span vectors with k-means segmentation. *ResearchGate*.

[Menezes et al. 2024] Menezes, R. P. B., Scotti, L., and Scotti, M. T. (2024). Aprendizado de máquina aplicado a qsar. *Química Nova*, 47(7):e–20240024.

[Mosquéra et al. 2024] Mosquéra, L. R., de Oliveira, M. N., dos Santos Martins, P. H., et al. (2024). Biofuel dynamics in brazil: Ethanol–gasoline price threshold analysis for consumer preference. *Energies*, 17(21):5265.

[Patil et al. 2024] Patil, R. V., Aggarwal, R., Poddar, G. M., Bhowmik, M., and Patil, M. K. (2024). Embedded integration strategy to image segmentation using canny edge and k-means algorithm. *International Journal of Intelligent Systems and Applications in Engineering*, 12(13s):1–8.

[Pavlopoulos et al. 2024] Pavlopoulos, I., Andreadis, S., and Vazirgiannis, M. (2024). Revisiting silhouette aggregation. *arXiv preprint arXiv:2401.05831*.

[Quintino et al. 2022] Quintino, D. D., Burnquist, H. L., and Ferreira, P. (2022). Relative prices of ethanol-gasoline in the major brazilian capitals: An analysis to support public policies. *Energies*, 15(13):4795.

[Rousseeuw 1987] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

[Shahapure 2020] Shahapure, K. R. (2020). Cluster quality analysis using silhouette score. *M.S. thesis*.

[Sinaga and Yang 2020] Sinaga, K. P. and Yang, M.-S. (2020). Unsupervised k-means clustering algorithm. *IEEE Access*, 8:80716–80727.

[Staszczak 2019] Staszczak, D. E. (2019). Recessions and a changing theoretical basis of the recoveries: a view from the state-corporation hegemonic stability theory. *Brazilian Journal of Political Economy*, 39(4):675–688.

[Wala et al. 2024] Wala, J., Herman, Umar, R., and Suwanti (2024). Heart disease clustering modeling using a combination of the k-means clustering algorithm and the elbow method. *Scientific Journal of Informatics*, 11(4):903–914.

Apêndice A: Dicionário de Dados

Dicionário de Dados das Colunas Geradas			
Atributo	**Descrição**	**Tipo**	**Categorias/Valores**
regiao	Região geográfica agregada do registro.	Char	norte, nordeste, centro-oeste, sudeste, sul
estado	Nome do estado onde ocorreu a coleta.	Char	alagoas, minas gerais, são paulo, etc.
municipio	Município onde ocorreu a coleta.	Char	arapiraca, serra, manaus, etc.
revenda	Nome da revenda de combustíveis.	Char	auto posto chaves ltda, posto ajuruteua ltda, auto posto kika ltda, etc.
cnpj_revenda	Número do CNPJ da revenda.	Char(14)	12345678912345
produto	Nome do produto comercializado.	Char	gasolina, etanol, diesel, etc.
data_coleta	Data em que o preço foi coletado.	Date	2024-04-29 (aaaa-mm-dd)
valor_venda	Preço de venda do produto na data da coleta (R$/unidade).	Float	Entre 2.63 e 9.79
unidade_medida	Unidade de medida do produto.	Char	R$ / litro
bandeira	Bandeira da revenda (marca).	Char	ipiranga, vibra, raizen etc
safra_coleta	Período/safra de referência associado à coleta.	Char	Entre 202401 e 202412
preco_medio_produto	Preço médio do produto no período base.	Float	Entre 4.07 e 6.12
preco_medio_produto_safra	Preço médio do produto por safra (mês).	Float	Entre 3.70 e 6.34
preco_medio_produto_safra_municipio	Preço médio do produto por safra e por município.	Float	Entre 2.81 e 8.02
preco_medio_produto_safra_estado	Preço médio do produto por safra e por estado.	Float	Entre 3.01 e 7.78
preco_medio_produto_safra_regiao	Preço médio do produto por safra e por região.	Float	Entre 3.43 e 6.91
var_preco_media_produto	Variação percentual do preço médio do produto.	Float	Entre -1.44 e 3.66
var_preco_media_produto_safra	Variação percentual do preço médio do produto por safra (mês).	Float	Entre -1.49 e 3.50
var_preco_media_produto_safra_municipio	Variação percentual do preço médio do produto por safra e por município.	Float	Entre -1.89 e 3.56
var_preco_media_produto_safra_estado	Variação percentual do preço médio do produto por safra e por estado.	Float	Entre -1.88 e 3.46
var_preco_media_produto_safra_regiao	Variação percentual do preço médio do produto por safra e por região.	Float	Entre -1.73 e 3.49
qnt_coletas_produto	Quantidade de coletas do produto.	Int	Entre 18941 e 233016
qnt_coletas_produto_safra	Quantidade de coletas do produto por safra (mês).	Int	Entre 1372 e 22735
qnt_coletas_produto_safra_municipio	Quantidade de coletas do produto por safra e por município.	Int	Entre 1 e 1035
qnt_coletas_produto_safra_estado	Quantidade de coletas do produto por safra e por estado.	Int	Entre 1 e 6054
qnt_coletas_produto_safra_regiao	Quantidade de coletas do produto por safra e por região.	Int	Entre 5 e 10639
qnt_produtos_revenda_produto	Quantidade de produtos comercializados pela revenda e por produto.	Int	Entre 1 e 56
qnt_produtos_revenda_safra	Quantidade de produtos comercializados pela revenda e por safra.	Int	Entre 1 e 6
qnt_produtos_revenda_safra_municipio	Quantidade de produtos da revenda por safra e por município.	Int	Entre 1 e 6
qnt_produtos_revenda_safra_estado	Quantidade de produtos da revenda por safra e por estado.	Int	Entre 1 e 6
qnt_produtos_revenda_safra_regiao	Quantidade de produtos da revenda por safra e por região.	Int	Entre 1 e 6
qnt_revendas_produto	Quantidade de revendas que comercializam o produto.	Int	Entre 18941 e 233016
qnt_revendas_safra	Quantidade de revendas por safra (mês).	Int	Entre 61686 e 88652
qnt_revendas_municipio	Quantidade de revendas por município.	Int	Entre 63 e 39481
qnt_revendas_estado	Quantidade de revendas por estado.	Int	Entre 2662 e 245555
qnt_revendas_regiao	Quantidade de revendas por região.	Int	Entre 59194 e 433910
preco_relativo_produto	Preço relativo (índice normalizado) do produto.	Float	Entre 0.64 e 1.80
preco_relativo_safra	Preço relativo (índice) por safra.	Float	Entre 0.69 e 1.78
preco_relativo_municipio	Preço relativo (índice) por município.	Float	Entre 0.68 e 1.70
preco_relativo_estado	Preço relativo (índice) por estado.	Float	Entre 0.71 e 1.67
preco_relativo_regiao	Preço relativo (índice) por região.	Float	Entre 0.72 e 1.61

Figura 5. *

Fonte: A autora (2025).