ISSN: 2317-0840

Avaliação de similaridade entre clusterizações hierárquica e particionada: Estudo com preços de combustíveis

Wylliam Eduardo Alves Silva[1†], Iêda Maria de Siqueira Bezerra[1] , Mayara Macedo da Mata[2]

[1]*Universidade Federal Rural de Pernambuco, Departamento de Estatística e Informática, Programa de Pós-Graduação em Biometria e Estatística Aplicada, Recife - Pernambuco, Brasil.*
[2]*Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, Programa de Pós Graduação em Química, Campina Grande - Paraíba, Brasil.*

Resumo: *O presente trabalho teve como objetivo comparar a similaridade entre os agrupamentos gerados pelos métodos hierárquico e particionado (k-means) na análise de preços de combustíveis, utilizando dados de gasolina comum e etanol no município de Campina Grande – PB em 2019. A análise de cluster, técnica estatística multivariada, foi aplicada para classificar postos de combustíveis em grupos homogêneos, considerando a proximidade dos preços. As análises foram realizadas no software R, demonstrando a viabilidade da aplicação dessas técnicas em dados reais. No método não hierárquico (k-means), a definição do número de grupos foi feita por meio do método Elbow, sendo identificadas diferenças significativas nas médias de preços entre os grupos. No método hierárquico, empregou-se a distância euclidiana e ligação completa, os clusters obtidos apresentaram estruturas semelhantes à do k-means. A comparação entre as duas abordagens revelou consistência na formação dos grupos e na similaridade estrutural dos clusters gerados, indicando que ambos os métodos produziram resultados convergentes para a segmentação dos postos por preço evidenciando a confiabilidade das técnicas utilizadas. Conclui-se que a análise de cluster é uma ferramenta eficaz para estudos de mercado de combustíveis, e que os métodos hierárquico e particionado, embora distintos em sua abordagem, geraram agrupamentos coerentes neste contexto.*

Palavras-chave: *Análise de cluster; Métodos hierárquicos e particionados; Preços de combustíveis; k-means; Distância euclidiana.*

Introdução

Postos de combustíveis são instalações que comercializam derivados de petróleo e biocombustíveis, como gasolina, diesel, etanol e gás natural, sendo regulados pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) (PETRÓLEO, G. N. E. B. AGÊNCIA NACIONAL do, 2017). As oscilações nos preços dos combustíveis, são fortementes influenciadas pela volatilidade do mercado internacional de petróleo, que geram efeitos significativos sobre o desempenho econômico e o consumo das famílias *(Dallal et al., 2024)*. Em particular, a gasolina e o etanol, combustíveis amplamente utilizados no Brasil, têm apresentado trajetórias de preços interligadas, influenciadas por fatores como custos de produção, tributos e concorrência local.

Nesse contexto, técnicas estatísticas multivariadas, como a análise de *clusters*, tornam-se ferramentas valiosas para identificar padrões de similaridade entre preços de combustíveis em diferentes estabelecimentos. Esta abordagem, classificada como aprendizado de máquina não supervisionado, agrupa observações com base em suas similaridades, maximizando a homogeneidade intraclusters e a heterogeneidade entre *clusters* (Silva, 2021). Esta técnica tem se mostrado versátil em distintas áreas do conhecimento, especialmente nas ciências econômicas

[†]Autor correspondente: wylliameduardo99@gmail.com
Manuscrito recebido em: 16/04/2025. Revisado em: 12/05/2025. Aceito em:19/05/2025.

e gerenciais, ao possibilitar a organização de grandes volumes de dados em grupos com características semelhantes, desse modo, facilitando a análise e interpretação de padrões relevantes (Ferreira et al., 2020).

Dentre as técnicas de agrupamento disponíveis, destacam-se os métodos hierárquicos e particionados, que se diferenciam principalmente pela estratégia adotada para formar os grupos ao longo do processo de análise (Pereira, 2023). Enquanto os métodos hierárquicos agrupam os dados com base em medidas de similaridade ou dissimilaridade, resultando em uma estrutura em forma de dendrograma, os métodos particionados, como o *k-means*, requerem a definição prévia do número de grupos e utilizam um processo iterativo para otimizar a alocação dos elementos a cada cluster (Oliveira et al., 2022). A escolha entre essas técnicas depende dos objetivos do estudo e da natureza dos dados, mas comparar sua eficácia em gerar agrupamentos consistentes é essencial para validar resultados.

Este trabalho busca avaliar a similaridade entre os agrupamentos formados pelos métodos hierárquico e particionado na análise de preços de gasolina comum e etanol em Campina Grande – PB. Para isto, aplicou-se a análise de *clusters* visando identificar postos com preços semelhantes, comparando a coerência dos grupos gerados por cada método.

Materiais e métodos

Os dados utilizados neste estudo foram coletados presencialmente pelo Fundo Municipal de Defesa dos Direitos Difusos (PROCON) de Campina Grande - PB, referentes aos preços de combustíveis no ano de 2019. O banco de dados compreende 57 postos de combustíveis distribuídos pela cidade, contendo as seguintes variáveis: identificação do posto, bandeira, bairro, preço da gasolina comum e preço do etanol. Após a coleta, realizou-se uma verificação minuciosa para identificar e tratar possíveis dados faltantes ou inconsistências no armazenamento, garantindo a qualidade dos dados antes das análises.

Todos os procedimentos estatísticos, cálculos e visualizações foram realizados no software R (R Core Team, 2024), utilizando pacotes especializados para análise de dados. Dentre os principais pacotes empregados, destacam-se: cluster (para análise de *cluster*) (Maechler et al., 2025), factoextra (para visualização de dados multivariados) (Kassambara; Mundt, 2020), ClustOfVar (para clusterização de variáveis) (Chavent et al., 2025) e ggdendro (para criação de dendrogramas) (de Vries; Ripley, 2024).

Análise de cluster

A análise de *cluster*, também conhecida como análise de agrupamentos, é um grupo de técnicas multivariadas cuja finalidade principal agregar objetos com base nas características que eles possuem (Hair et al., 2009). Os objetos, em cada grupo, tendem a ser semelhante entre si, mas diferentes de objetos em outros grupos, (Malhotra, 2019).

Conforme Falqueto e Cezar (2022), o processo de análise de *cluster* envolve diversas etapas fundamentais, que incluem a definição da amostra a ser agrupada, a seleção das variáveis mais relevantes para representar as características dos indivíduos e a escolha de um método de agrupamento, seja por agregação ou partição, garantindo a avaliação da qualidade e coerência dos agrupamentos formados.

Neste trabalho, a determinação do número de *clusters* foi realizada por meio do método do *Elbow* (cotovelo), que geralmente é considerado um bom indicador da quantidade apropriada de agrupamentos.

O método do cotovelo representa uma das abordagens mais utilizadas tradicionalmente para estimar o número ideal de *clusters* em uma base de dados, sendo amplamente utilizado pela simplicidade de sua aplicação, apesar de sua sensibilidade a avaliação gráfica (Alves et al., 2024).

Medidas de semelhança e distância

De acordo com Paz (2024), existem duas abordagens principais para quantificar a relação entre objetos em uma análise: as medidas de similaridade, que expressam o grau de correspondência direta entre elementos, sendo os valores mais altos indicativos de maior semelhança, e as medidas de dissimilaridade ou distância, nas quais valores mais elevados refletem maior divergência entre os itens comparados.

Coeficientes de correlação

As medidas de similaridade possuem um papel crucial na análise de agrupamentos, possibilitando quantificar o grau de associação entre objetos. Dentre estas medidas, o coeficiente de correlação de Pearson se destaca por sua capacidade de capturar relações lineares entre variáveis, sendo comumente aplicado em diversas áreas, devido à sua interpretação intuitiva e aplicação simples (Albuquerque et al., 2022). Embora o coeficiente de correlação seja comumente atribuído a Karl Pearson, sua origem remete aos estudos de Francis Galton, cujas ideias sobre regressão e correlação influenciaram diretamente o trabalho de Pearson (Chattamvelli, 2024).

Medidas de distância

Grande parte das técnicas de agrupamento depende do cálculo de uma medida que quantifique o grau de separação entre os objetos, geralmente representada por funções de distância ou métricas específicas (Gao et al., 2023). Diversas funções podem ser utilizadas como medidas de distância em análises de agrupamento (Buccianti; Gozzi, 2023). Algumas distâncias estão enquadradas na definição de medidas de similaridade, como: Distância euclidiana, Distância Euclidiana Quadrática, Manhattan, Distância de Minkowski e Mahalanobis (Crispim et al., 2019).

Métodos hierárquicos

Este procedimento organiza os dados em uma estrutura hierárquica, agrupando os elementos conforme suas semelhanças (Ran et al., 2023). Os métodos hierárquicos podem ser classificados em dois tipos principais: os aglomerativos, que constroem os agrupamentos por meio de junções progressivas entre os elementos, e os divisivos, que os formam a partir de sucessivas separações (Wang et al., 2023). O método aglomerativo inclui uma variedade de técnicas amplamente utilizadas na prática, (Abushilah; Abbas, 2023): *single linkage*, *complete linkage*, *average linkage*, *median linkage*, método do centroide, método de *Ward*. De acordo com Cabezas et al. (2023), nos métodos hierárquicos, é comum utilizar um dendrograma ou diagrama em forma de árvore, para ilustrar os agrupamentos. Neste tipo de representação, os ramos indicam os elementos individuais, enquanto a raiz simboliza o agrupamento total.

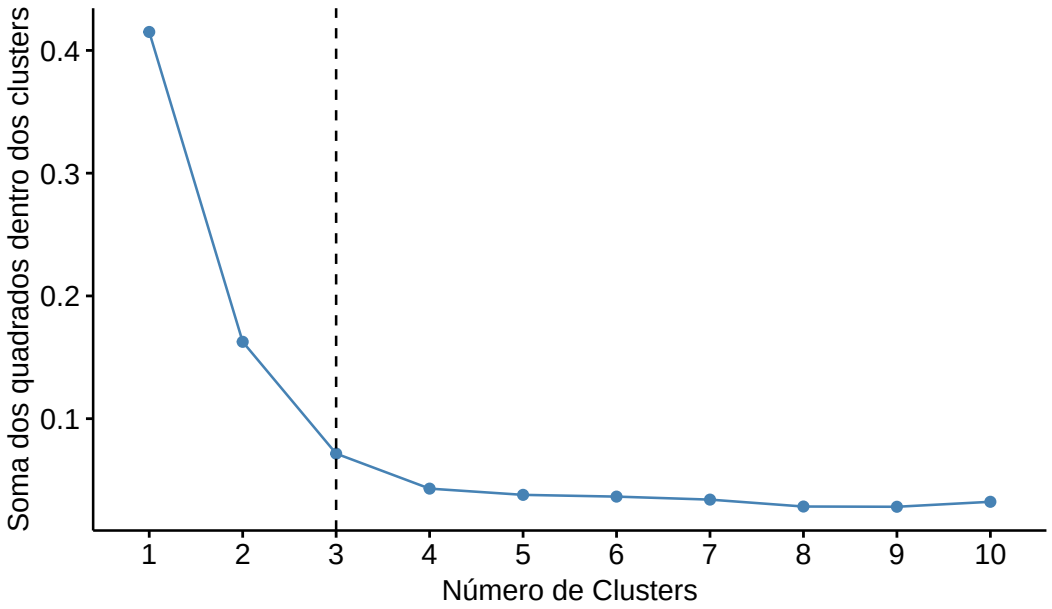
Resultados e discussões

Primeiramente, foram aplicadas as técnicas de análise de *cluster*, tanto a abordagem não hierárquica (método *k-means*) quanto a hierárquica (com ligação completa e distância euclidiana), para identificar agrupamentos naturais entre os postos de combustível com base nos preços de gasolina e etanol. Após a definição dos *clusters* por cada método, procedeu-se com uma análise descritiva detalhada para os preços em cada grupo formado.

Método não hierárquico

Obtendo o número de *cluster*s através de método de *Elbow* (cotovelo), que se observa através do *scree plo*t, por esse método a localização de uma curva (cotovelo) no gráfico é geralmente considerada como um indicador do número apropriado de *clusters*.

Figura 1: Determinação do número ideal de *clusters* com base no método de Elbow *scree plot*

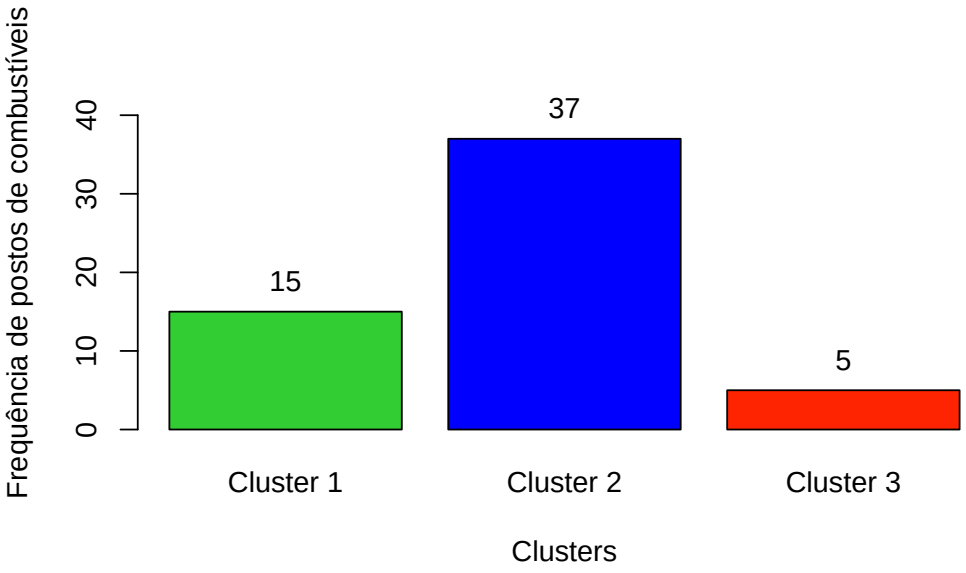


Fonte: dos autores (2025).

Observando o *scree plot* podemos verificar uma desaceleração no decaimento, assim o número ideal de *clusters* segundo o método de *Elbow* são 3 (três)

Utilizando o método *K-means*, é possível observar a distribuição dos postos de combustíveis entre os diferentes *clusters*. Essa segmentação pode ser visualizada na Figura 2.

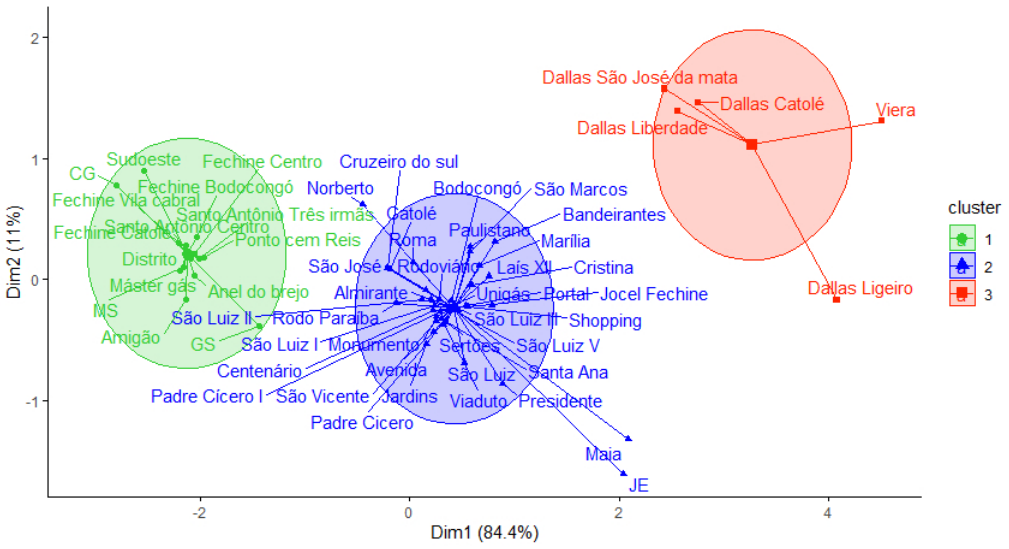
Figura 2: Distribuição da quantidade de postos de combustíveis por *cluster* identificados



Fonte: dos autores (2025).

A Figura 3 apresenta a localização dos postos de combustíveis organizados de acordo com os *clusters* formados pelo método *K-means*. Cada cor representa um grupo distinto.

Figura 3: Localização dos postos nos *clusters*

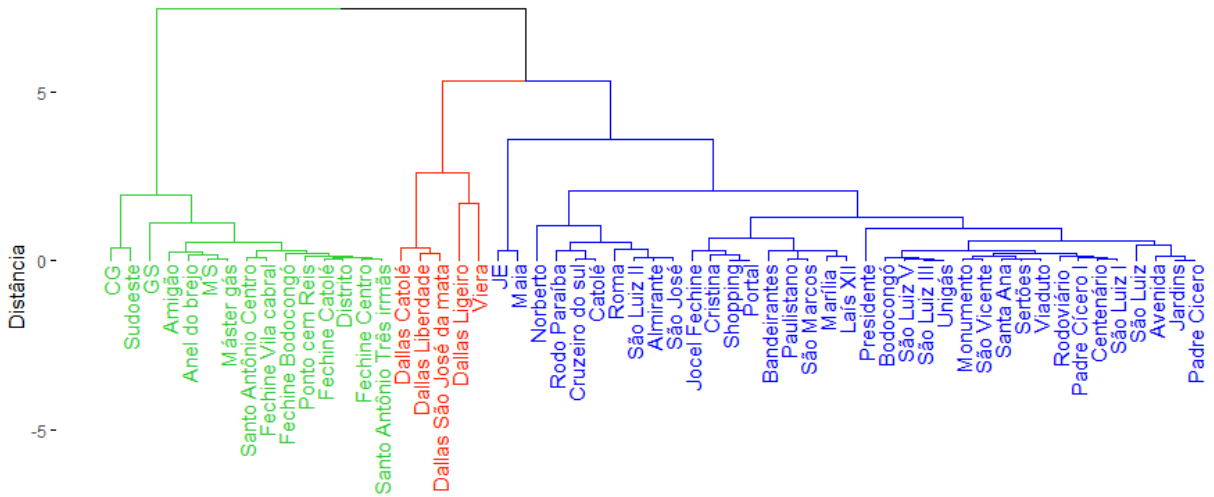


Fonte: dos autores (2025).

Método hierárquico

A Figura 4 apresenta o dendrograma evidenciando os *clusters* formados pelo método *complete linkage* (ligação completa), seguindo o mesmo critério utilizado para obter o número de *cluster* no método não hierárquico, ou seja, formando 3 *clusters*.

Figura 4: Dendrograma



Fonte: dos autores (2025).

Estatísticas descritivas por cluster

No intuito de se fazer um levantamento sobre algumas características importantes, a Tabela 1 apresenta algumas estatísticas descritivas obtidas para os *clusters* em relação ao preço da gasolina comum.

Tabela 1: Sumário das estatísticas para gasolina comum por *clusters*

Estatísticas	cluster 1	cluster 2	cluster 3
Mínimo	4,257	4,294	4,318
Mediana	4,283	4,327	4,326
Média	4,282	4,328	4,343
Desvio padrão	0,012	0,017	0.030
CV	0,003	0,004	0.007
Máximo	4,308	4,388	4,386

Fonte: dos autores (2025).

Observando as estatísticas para cada *cluster* o que chama atenção é o *cluster* 3 que tem um total de 5 postos de combustíveis e teve a maior média de preço, indicando que esses 5 postos vendiam a gasolina comum com os preços mais elevados durante o ano de 2019. Os valores da média indicam que os postos com os melhores preços são os do *cluster* 1: um total de 15 postos, pois apresenta a menor média de preço para a gasolina comum, o *cluster* 2 com um total de 37 tem a média entre o *cluster* 1 e 3.

A Tabela 2 apresenta algumas estatísticas descritivas obtidas para os *clusters* em relação ao preço do etanol, no intuito de se fazer um levantamento sobre algumas características importantes.

Tabela 2: Sumário das estatísticas para etanol por *clusters*

Estatísticas	cluster 1	cluster 2	cluster 3
Mínimo	3,209	3,268	3,478
Mediana	3,235	3,299	3,499
Média	3,232	3,312	3,523
Desvio padrão	0,013	0,031	0,068
CV	0,004	0,009	0,019
Máximo	3,257	3,391	3,641

Fonte: dos autores (2025).

Portanto, o que foi observado para a gasolina comum acontece com o etanol, o *cluster* 3 tem o maior preço médio para o etanol, além dos maiores valores para todas as estatísticas apresentadas, como para a gasolina comum o menor preço médio se encontra no *cluster* 1.

Conclusão

Os resultados indicaram que existem 3 *clusters* alocando os postos com os preços mais similares. O método hierárquico utilizou a distância euclidiana e a ligação completa, no não hierárquico utilizou o método *k-means*, ou seja, isso indica que a análise de *cluster* foi bem empregada neste estudo, pois em ambos os métodos não houve mudança dos postos de *clusters*.

Como destacado por Silva (2021), acredita-se que o seguinte ponto é interessante de ser explorado, o Fundo Municipal de Defesa dos Direitos Difusos (PROCON) de Campina Grande - PB poderia implementar a análise de *cluster* nas suas pesquisas de preços de combustíveis que é realizada mensalmente, ajudando a população a identificar onde se encontra os postos com os preços mais acessíveis, assim, fazendo-os economizar com os gastos no abastecimento da gasolina comum e do etanol.

Referências

ABUSHILAH, S. F.; ABBAS, R. H. Performance evaluation of some clustering algorithms under different validity indices. **Mathematical Modelling of Engineering Problems**, v. 10, n. 4, p. 1271–1280, 2023.

ALBUQUERQUE, M. A. de; NASCIMENTO, E. R. do; BARROS, K. N. N. de O.; BARROS, P. S. N. Comparison between similarity coefficients with application in forest sciences. **Research, Society and Development**, v. 11, n. 2, p. e48511226046–e48511226046, 2022.

ALVES, K. A.; ALBUQUERQUE, O. de S.; SOUSA, A. L. de; JÚNIOR, G. de M. Análise de dados dos planos de desenvolvimento institucional do instituto federal do pará (2009-2023) utilizando o algoritmo de aprendizado de máquina k-means. **Cuadernos de Educación y Desarrollo**, v. 16, n. 11, p. e6259–e6259, 2024.

BUCCIANTI, A.; GOZZI, C. Cluster analysis and classification. In: **Encyclopedia of Mathematical Geosciences**. [S.l.]: Springer, 2023. p. 127–133.

CABEZAS, L. M.; IZBICKI, R.; STERN, R. B. Hierarchical clustering: Visualization, feature importance and model selection. **Applied Soft Computing**, Elsevier, v. 141, p. 110303, 2023.

CHATTAMVELLI, R. Measures of association. In: **Correlation in Engineering and the Applied Sciences: Applications in R**. [S.l.]: Springer, 2024. p. 1–54.

CHAVENT, M.; KUENTZ, V.; LIQUET, B.; SARACCO, J. Clustofvar: An r package for the clustering of variables (version 1.2). 2025. Disponível em: ⟨https://CRAN.R-project.org/package=ClustOfVar⟩.

CRISPIM, D. L.; FERNANDES, L. L.; ALBUQUERQUE, R. L. d. O. Aplicação de técnica estatística multivariada em indicadores de sustentabilidade nos municípios do marajó-pa. **Revista Principia**, v. 1, n. 46, p. 145–154, 2019.

DALLAL, G. F. C. et al. Flutuações no preço do petróleo e seus impactos em indicadores econômicos nacionais. Florianópolis, SC., 2024. Disponível em: ⟨https://repositorio.ufsc.br/handle/123456789/261480⟩.

de Vries, A.; RIPLEY, B. D. **ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'**. [S.l.], 2024. R package version 0.2.0. Disponível em: ⟨https://andrie.github.io/ggdendro/⟩.

FALQUETO, A. A.; CEZAR, L. C. SegmentaÇÃo via machine learning: Proposta de clusterizaÇÃo de consumidores do e-commerce de uma empresa multinacional do varejo esportivo. **HOLOS**, v. 4, dez. 2022. Disponível em: ⟨https://www2.ifrn.edu.br/ojs/index.php/HOLOS/article/view/12032⟩.

FERREIRA, R.; PAIM, F. d. P.; RODRIGUES, V.; CASTRO, G.; RODRIGUES, U. V. G. S. Análise de cluster não supervisionado em r: agrupamento hierárquico. 2020. Disponível em: ⟨https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1126478/1/5360.pdf⟩.

GAO, C. X.; DWYER, D.; ZHU, Y.; SMITH, C. L.; DU, L.; FILIA, K. M.; BAYER, J.; MENSSINK, J. M.; WANG, T.; BERGMEIR, C. et al. An overview of clustering methods with guidelines for application in mental health research. **Psychiatry Research**, Elsevier, v. 327, p. 115265, 2023.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise multivariada de dados**. [S.l.]: Bookman editora, 2009.

KASSAMBARA, A.; MUNDT, F. **factoextra: Extract and Visualize the Results of Multivariate Data Analyses**. [S.l.], 2020. R package version 1.0.7. Disponível em: ⟨https://CRAN.R-project.org/package=factoextra⟩.

MAECHLER, M.; ROUSSEEUW, P.; STRUYF, A.; HUBERT, M.; HORNIK, K. **cluster: Cluster Analysis Basics and Extensions**. [S.l.], 2025. R package version 2.1.8.1 — For new features, see the 'NEWS' and the 'Changelog' file in the package source). Disponível em: ⟨https://CRAN.R-project.org/package=cluster⟩.

MALHOTRA, N. K. **Pesquisa de Marketing-: uma orientação aplicada**. [S.l.]: Bookman Editora, 2019.

OLIVEIRA, P. L. S. de; RODRIGUES, R. L.; RAMOS, J. L. C.; SILVA, J. C. S. Identificação de pesquisas e análise de algoritmos de clusterização para a descoberta de perfis de engajamento. **Revista Brasileira de Informática na Educação**, v. 30, p. 01–19, 2022.

PAZ, H. O. d. Método de agrupamento multinível para dados mistos. Universidade Federal da Bahia, 2024. Disponível em: ⟨https://repositorio.ufba.br/handle/ri/40414⟩.

PEREIRA, L. G. Clusterização como técnica de apoio à decisão para um marketplace eletrônico logístico. 2023. Disponível em: ⟨https://repositorio.unifei.edu.br/jspui/handle/123456789/3965⟩.

PETRÓLEO, G. N. E. B. AGÊNCIA NACIONAL do. **Cartilha do posto revendedor de combustíveis**. 2017. Disponível em: ⟨https://www.gov.br/anp/pt-br/centrais-de-conteudo/publicacoes/cartilhas-e-guias/arq/cartilhapostorevendedor6ed.pdf⟩.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2024. Disponível em: ⟨https://www.R-project.org/⟩.

RAN, X.; XI, Y.; LU, Y.; WANG, X.; LU, Z. Comprehensive survey on hierarchical clustering algorithms and the recent developments. **Artificial Intelligence Review**, Springer, v. 56, n. 8, p. 8219–8264, 2023.

SILVA, W. E. A. **Análise de cluster aplicada aos dados de preços de combustíveis na cidade de Campina Grande - PB**. 31 p. — Universidade Estadual da Paraíba, Campina Grande, 2021. Trabalho de Conclusão de Curso (Graduação em Estatística). Disponível em: ⟨http://dspace.bc.uepb.edu.br/jspui/handle/123456789/25640⟩.

WANG, F.; ZHOU, G.; XIE, J.; FU, B.; YOU, H.; CHEN, J.; SHI, X.; ZHOU, B. An automatic hierarchical clustering method for the lidar point cloud segmentation of buildings via shape classification and outliers reassignment. **Remote Sensing**, MDPI, v. 15, n. 9, p. 2432, 2023.