Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Allysson de Souza Pereira

Clusterização de Bisnagas: Um Enfoque em Eficiência Operacional na Indústria de Cosméticos

Curitiba
2025

Allysson de Souza Pereira

Clusterização de Bisnagas: Um Enfoque em Eficiência Operacional na Indústria de Cosméticos

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Cesar Taconeli
Coorientador: Jonh Lemos

Curitiba
2025

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Clusterização de Bisnagas: Um Enfoque em Eficiência Operacional na Indústria de Cosméticos

Allysson de Souza Pereira[1]
Cesar Taconeli[2]
Jonh lemos[3]

Resumo

No contexto da Indústria 4.0, a indústria de cosméticos recorre à Inteligência Artificial e à cultura *Data Driven* como diferenciais estratégicos para antever tendências e otimizar recursos. O objetivo central deste artigo é aplicar técnicas de aprendizagem de máquina não supervisionada — especificamente K-Means e Clusterização Hierárquica — para mapear e compreender os perfis de grupos em um portfólio com mais de 2.000 SKUs de bisnagas, utilizando suas características físicas e dimensionais. A partir da identificação de famílias homogêneas de produtos, busca-se detectar oportunidades para maximizar a performance do processo fabril, otimizando o sequenciamento da produção e a eficiência global (OEE), superando as limitações da categorização tradicional. A metodologia adotada foi quantitativa e exploratória, aplicando a transformação Box-Cox na engenharia de *features* e simulando 8 cenários de modelagem para testar a robustez do agrupamento em conjuntos de dados distintos. As análises comparativas demonstraram que a mitigação da redundância dimensional nos dados de entrada maximizou a coesão dos grupos. O modelo K-Means com $K = 65$ (Cenário C8) foi selecionado como a solução de melhor balanço estratégico e estatístico (Silhueta: 0,5733; DBI: 0,5240). Conclui-se que a caracterização robusta em 65 famílias fornece subsídios acionáveis para a gestão industrial elevar a padronização e a utilização dos ativos de fabricação.

Palavras-chave: Indústria 4.0; Clusterização; K-Means; Clusterização Hierárquica Aglomerativa.

Abstract

In the context of Industry 4.0, the cosmetics industry turns to Artificial Intelligence and a *Data Driven* culture as strategic differentiators to forecast trends and optimize resources. The central objective of this paper is to apply unsupervised machine learning techniques—specifically K-Means and Hierarchical Clustering—to map and understand group profiles within a tube portfolio containing over 2,000 SKUs, utilizing their physical and dimensional characteristics. Based on the identification of homogeneous product families, the study seeks to detect opportunities to maximize manufacturing process performance, optimizing production sequencing and Overall Equipment Effectiveness (OEE), overcoming the limitations of traditional categorization. The methodology adopted was quantitative and exploratory, applying Box-Cox transformation for feature engineering and simulating 8 modeling scenarios to test clustering robustness across distinct datasets. Comparative analyses demonstrated that mitigating dimensional redundancy in the input data maximized group cohesion. The K-Means model with $K = 65$ (Scenario C8) was selected as the solution with the best strategic and statistical balance (Silhouette: 0.5733; DBI: 0.5240). It is concluded that the robust characterization into 65 families provides actionable insights for industrial management to enhance standardization and the utilization of manufacturing assets.

Keywords: Industry 4.0; Clustering; K-Means; Agglomerative Hierarchical Clustering.

1 Introdução

O mercado de higiene pessoal e cosméticos no Brasil é um dos mais dinâmicos do mundo, com o setor projetando um crescimento anual de 7% até 2027, evidenciando sua robustez e sua importância estratégica global [1]. As empresas deste mercado caracterizam-se por serem extremamente ágeis em seus lançamentos, a fim de antever tendências em um mercado cada vez mais competitivo [2], e suprir as expectativas de um consumidor que busca uma equação custo-benefício mais vantajosa. O reflexo disso, segundo a Associação Brasileira da Indústria de Higiene Pessoal, Perfumaria e Cosméticos [3], é uma indústria inovadora, ágil e focada em fatores qualitativos e quantitativos.

Tal agilidade reflete diretamente no número de produtos lançados: em 2021, a indústria brasileira lançou

[1]Aluno do programa de Especialização em Data Science & Big Data, allyssonsouza1@hotmail.com.
[2]Professor do Departamento de Estatística - DEST/UFPR.
[3]Doutorando do Programa de Pós-Graduação em Engenharia Elétrica e de Computação - LSD/UFBA.

7.368 produtos, ultrapassando a China e se tornando a segunda maior potência mundial em lançamentos. Com o advento da indústria 4.0 e o avanço da tecnologia, as organizações se tornaram mais inovadoras e disruptivas nos seus processos e tarefas, buscando minimizar custos, mitigar erros e aumentar a competitividade [2].

Neste cenário de intensa transformação, o desenvolvimento da Inteligência Artificial (IA) e das práticas de Ciência de Dados se tornou um diferencial estratégico. Empresas que adotam uma cultura Data Driven[2] utilizam essas tecnologias para otimizar o modelo de tomada de decisão, tornando-o mais assertivo e eficiente [4]. Uma das formas mais poderosas de aplicação é a implementação de modelos de aprendizagem de máquina.

Aprendizagem de máquina é uma subárea da Inteligência Artificial, que utiliza modelos matemáticos para inferir aprendizado baseado em exemplos [5]. A aprendizagem de máquina se subdivide em categorias, sendo o aprendizado não supervisionado fundamental para a clusterização. Segundo [6], clusterização consiste em subdividir uma população heterogênea em subgrupos mais coesos, sem rótulos pré definidos, identificando grupos por meio de suas semelhanças.

Segundo Metz [7], a clusterização é uma ferramenta essencial para a detecção e segmentação de características, sendo um método cujos resultados dependem diretamente da escolha de parâmetros como as medidas de similaridade utilizadas.

A aplicação dessa técnica é vasta:[6] a utilizou para a predição de evasão escolar,[4] usou o método de clusterização para definição de um sistema de recomendação baseando-se no perfil de compra de cada grupo e [8] demonstraram o potencial do K-Means na análise e interpretação de vastos volumes de dados provenientes de imagens de satélite, destacando sua aplicação em áreas como monitoramento ambiental e agricultura de precisão.

Diante do forte crescimento na aplicação das técnicas de aprendizagem de máquina em diversas áreas, como saúde, transportes e automobilística [4], o mercado de cosméticos também se destaca. Neste cenário, este trabalho vem de encontro com essa demanda, propondo uma solução eficiente e acessível para a categorização de SKUs[2] em famílias de bisnagas, grupos que deveriam compartilhar especificações técnicas para otimizar o processo produtivo. Atualmente, esse processo é executado de forma manual e subjetiva, o que o torna suscetível a erros, resultando em decisões baseadas na experiência individual e desprovidas de critérios objetivos. Essas oportunidades de melhoria são claramente visualizadas na eficiência e gestão de rotina do setor operacional.

A segmentação não padronizada de SKUs com características físicas e dimensionais distintas pode levar a alocação subótimas de recursos e a gargalos no processo produtivo, como o envase, impactando a performance e a flexibilidade produtiva. Portanto, a oportunidade reside na substituição dessa abordagem subjetiva por um método capaz de realizar a segmentação automática de SKUs[2] em famílias de forma precisa, ágil e padronizada. Este método deve garantir maior flexibilidade de roteiro de produção, minimizar a necessidade de troca de ferramentais e, consequentemente, aumentar a eficiência global da produção (*Overall Equipment Effectiveness* - [3]OEE).

Diante deste panorama, o objetivo geral deste artigo é aplicar técnicas de aprendizagem de máquina não supervisionado — especificamente os algoritmos K-Means e Clusterização Hierárquica — para a definição de grupos homogêneos em um portfólio de bisnagas. O estudo utilizará como variáveis as características físicas e dimensionais dos SKUs[2]. A finalidade é gerar informações estratégicas que otimizem o sequenciamento produtivo, a padronização e a melhor utilização de ativos e ferramentais de fabricação.

As contribuições deste artigo são:

- Desenvolver um modelo de clusterização robusto para o portfólio de bisnagas, capaz de agrupar os SKUs com base em suas características físicas e dimensionais.
- Realizar uma análise detalhada usando os métodos K-Means e Clusterização Hierárquica, incluindo a correta definição do número de clusters (parâmetro K) com base em critérios estatísticos e contextuais de negócio.

[1] **Data Driven:** O termo *Data Driven* significa, em português, "Orientado por Dados". Isso significa que uma empresa que possui uma cultura *Data Driven* baseia a maior parte dos seus processos e ações na coleta e análise de dados, visando a tomada de decisão assertiva. Disponível em: https://www.alura.com.br/artigos/data-driven?srsltid=AfmBOoqcvIVTyHolcML3rGJlqqe7SRR1AhDAbipfBDx7qOgGPEpJIXIa.
[2] **SKU:** A sigla vem do termo em inglês *Stock Keeping Unit*, ou Unidade de Manutenção de Estoque. Disponível em: https://venda.amazon.com.br/sellerblog/o-que-e-sku-do-produto-e-qual-a-importancia-de-utilizar-esse-codigo.
[3] **OEE:** Utilizado na indústria, o *Overall Equipment Effectiveness* aponta o nível de eficiência de um equipamento. Disponível em: https://www.totvs.com/blog/gestao-industrial/oee/.

2 Materiais e Métodos

O estudo é caracterizado como uma pesquisa aplicada, dado seu objetivo prático em desenvolver um modelo capaz de prover um agrupamento ótimo de produtos com base em características dimensionais e processuais. A abordagem é quantitativa e de natureza exploratória, utilizando técnicas de aprendizagem de máquina não supervisionada para identificar os padrões e estruturas em dados não rotulados. A metodologia segue etapas que abrangem coleta de dados, pré-processamento, a modelagem por agrupamento e validação, baseando-se em aspectos científicos e regras de negócio, conforme demonstrado no fluxograma da Figura 1.

2.1 Conjunto de Dados

O conjunto de dados foi extraído do sistema Systems, Applications & Products (SAP) e de tabelas construídas durante o estudo em uma empresa de grande porte do setor de cosméticos e higiene pessoal no Brasil. Devido

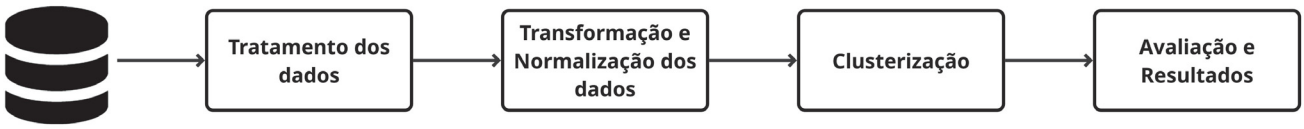

Figura 1: Fluxograma de etapas empregadas no trabalho. Fonte: O autor

ao sigilo empresarial, o nome da organização e os valores absolutos das variáveis são omitidos.

O objeto do estudo é o portfólio de mais de 2.000 produtos da categoria de bisnagas de diversos modelos, cores, tamanhos e volumetrias. A escolha deste segmento se justifica pela sua alta representatividade no volume de produção e pela grande oportunidade de automatizar o processo de definição de famílias. A clusterização desses produtos visa trazer celeridade e grandes impactos nas áreas incumbidas pela gestão do portfólio, especialmente na padronização de ferramental e no sequenciamento produtivo.

A seleção das variáveis para o modelo de clusterização foi um passo metodológico crucial. As variáveis foram elencadas visando prioritariamente sua relevância ao processo fabril, com o objetivo de obter uma melhor segmentação dos produtos desta categoria e identificar oportunidades de sequenciamento, ganho de setup (processo de preparação de um equipamento para iniciar uma nova produção) e padronização de ferramental de fabricação. A estrutura do conjunto de dados utilizado no modelo é apresentada na Tabela 1.

Tabela 1: Dados fictícios para representação do conjunto de dados.

diametro	altura	comprimento	peso	tp_selagem
13,00	36,22	23,23	120,00	tipo_1
5,00	15,55	8,39	90,00	tipo_2
34,00	17,00	25,33	100,00	tipo_3

Fonte: O autor (2025)

2.2 Limpeza e Preparo dos Dados

A preparação do conjunto de dados é uma fase crítica e fundamental em qualquer projeto de Ciência de Dados, sendo determinante para garantir a interpretabilidade e a confiabilidade dos resultados dos métodos. Todas as etapas de pré-processamento, engenharia de atributos e modelagem foram executadas no ambiente *Python*. A escolha do Python se justifica pela sua versatilidade, dinamismo e integração com diversas bibliotecas bem estabelecidas, como Pandas, *NumPy e Scikit-Learn*, oferecendo um robusto ferramental para a Ciência de Dados [9]. O conjunto de dados fornecido necessitou da aplicação de técnicas de transformação e normalização para que as variáveis atingissem o estado ideal para a modelagem.

A etapa inicial visou a higienização e a estruturação dos dados, garantindo sua integridade e formato para o processamento subsequente:

- **Carregamento e Junção dos Dados:** O processo foi iniciado pelo carregamento de diversas fontes de dados. Em seguida, foi realizada a junção, coletando as variáveis de interesse de cada base e formando um único conjunto de dados principal. Para isso, foi utilizada a função *merge* da biblioteca Pandas, definindo o código único de cada SKU como chave primária para garantir a unicidade e a consistência dos registros.
- **Tratamento de Tipos e Nomenclaturas:** Realizou-se a conversão dos tipos de dados, garantindo que as colunas numéricas fossem tratadas como *Float*. Houve também a padronização das nomenclaturas na coluna categórica de tp_selagem, consolidando múltiplas descrições textuais em categorias únicas e bem definidas.
- **Tratamento de Dados Inválidos e Outliers:** Por fim, foi realizado o tratamento de dados inconsistentes. Foram identificados e removidos registros com valores contextualmente impossíveis para as dimensões (a critério de exemplo, valores de altura igual a 0), além de categorias que representavam ruídos conceituais (*outliers*), garantindo que apenas SKUs válidos fossem considerados na modelagem exploratória.

2.2.1 Engenharia de Atributos

As variáveis dimensionais e de processo do portfólio apresentam duas características que influenciam diretamente a escolha do algoritmo: multimodalidade (presença de múltiplos picos na distribuição, como na variável altura) e assimetria, a qual persistiu mesmo após o tratamento de outliers. Tais distribuições podem comprometer a estabilidade de algoritmos baseados em distância. A visualização do perfil de distribuição das variáveis numéricas, antes das transformações, é ilustrada na Figura 2:

2.2.2 Transformação de Distribuição: Box-Cox

Para mitigar a forte assimetria observada, optou-se pela aplicação da Transformação Box-Cox [10], uma técnica paramétrica recomendada para estabilizar a variância e aproximar a distribuição dos dados de uma curva normal, pré-requisito crucial para modelos de clusterização baseados em distância. A transformação foi aplicada em

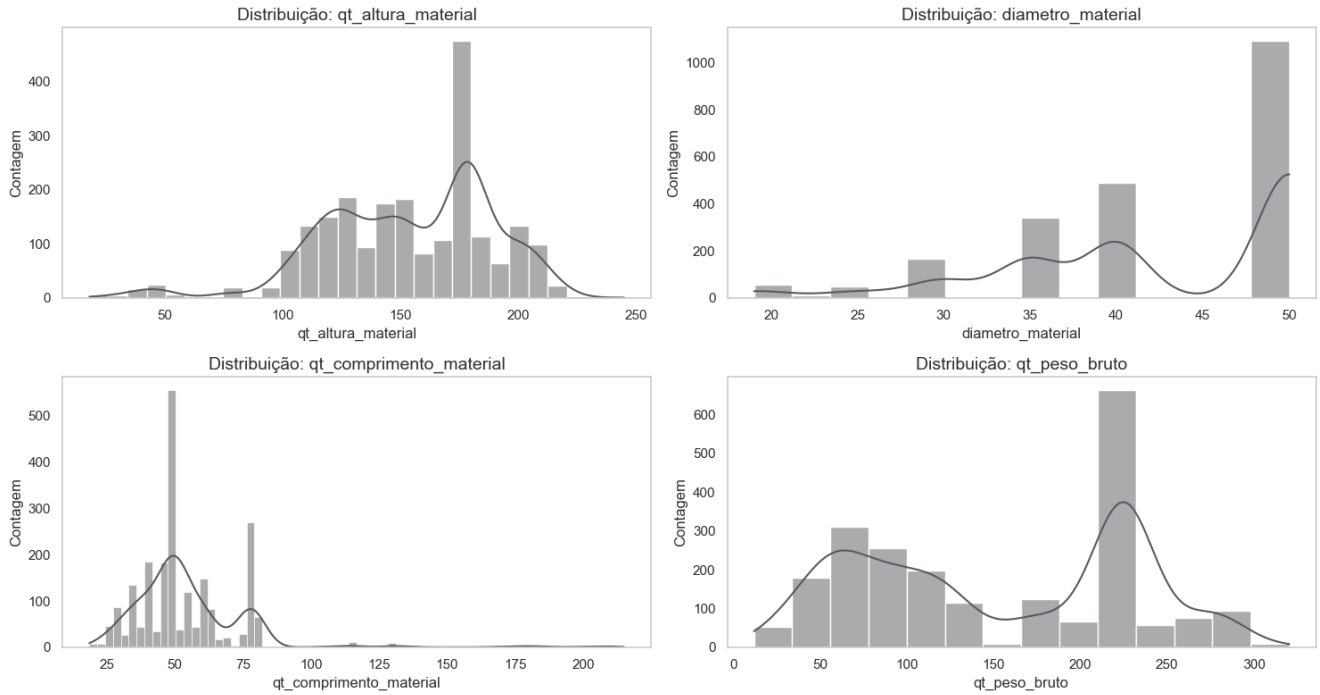

Figura 2: Representação do perfil das variáveis antes de passarem pelo processo de engenharia de atributos. Fonte: O autor(2025)

todas as variáveis numéricas: diâmetro, altura, comprimento e peso. Formalmente, a transformação é definida como:

$$y_i(\lambda) = \begin{cases} \frac{x_i^{\lambda}-1}{\lambda} & \text{se } \lambda \neq 0 \\ \ln(x_i) & \text{se } \lambda = 0 \end{cases} \quad (1)$$

onde:

- $\mathbf{y}_i(\lambda)$: Valor transformado.
- $\mathbf{x}_i$: original value.
- λ: Parâmetro de transformação (determinado pelo log-*likelihood* ótimo).
- $\ln(\mathbf{x}_i)$: Logaritmo Natural, aplicado quando $\lambda = 0$.

2.2.3 *One-Hot Encoding* de Variáveis Categóricas

A variável categórica de selagem, crucial para a distinção das famílias de bisnagas, foi submetida ao processo de codificação utilizando o método One-Hot Encoding [11]. Essa técnica converte a variável qualitativa em múltiplas colunas binárias (0 ou 1), onde cada nova coluna representa uma categoria de selagem específica (por exemplo, tipo_1, tipo_2). Este procedimento permite que a informação qualitativa seja processada de forma quantitativa pelos algoritmos de agrupamento baseados em distância.

2.2.4 Normalização dos dados

Uma vez transformadas as colunas, torna-se necessário a aplicação da normalização, visto que as colunas numéricas ainda possuíam diferentes escalas. O método utilizado para a resolução deste problema foi o *StandardScaler*. Esta técnica padroniza todas as colunas, formatando-as para que apresentem média zero ($\mu = 0$) e desvio-padrão unitário ($\sigma = 1$). A normalização é indispensável para o K-Means e a Clusterização Hierárquica, pois garante que as variáveis tenham pesos igualitários na definição dos clusters e no cálculo da distância euclidiana.

$$Z = \frac{x - \mu}{\sigma} \quad (2)$$

onde:

- Z: é o valor padronizado (*Z-score)*.
- x: é o valor original da variável.
- μ: é a média dos valores da variável no conjunto de dados.
- σ: é o desvio padrão dos valores da variável no conjunto de dados.

2.3 Métodos Empregados

Esta seção detalha os métodos de clusterização selecionados para a identificação de *clusters* homogêneos no portfólio de bisnagas. A escolha recaiu sobre os métodos K-Means e a clusterização Hierárquica Aglomerativa.

2.3.1 Método K-Means

O K-Means é um método de agrupamento particional baseado em centroides, fundamental para dividir um conjunto de dados não rotulados em um número fixo de K clusters. O objetivo principal é que cada cluster compartilhe características comuns, representadas pelo seu centróide.

Formalmente, o conjunto de dados de entrada pode ser representado pela matriz $\mathbf{X}$ (Figura 3), onde N é o número de observações e D é o número de variáveis [12]:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix}$$

Figura 3: Representação da Matriz de Dados ($\mathbf{X}$). A matriz X representa o conjunto de dados de entrada, onde cada linha é uma observação (SKU) e cada coluna é uma variável dimensional ou de processo.

A intenção do K-Means é minimizar a função de custo, denominada Inércia ou soma dos quadrados intra-cluster (*Within-Cluster Sum of Squares*). Esta função (Eq. 3) mede a coesão interna dos clusters, quantificando a soma das distâncias quadráticas (geralmente Euclidianas) entre cada amostra ($\mathbf{x}_i$) e o centróide ($\boldsymbol{\mu}_j$) de seu respectivo grupo:

$$\underset{C}{\operatorname{argmin}} \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in C_j} ||\mathbf{x}_i - \boldsymbol{\mu}_j||^2 \tag{3}$$

onde:

- $\underset{C}{\operatorname{argmin}}$: Instrução matemática de minimização da expressão seguinte em relação aos conjuntos de clusters (C).
- $\sum_{j=1}^{k}$: Somatório sobre todos os k *clusters* do modelo, indexados por j.
- $\sum_{\mathbf{x}_i \in C_j}$: Somatório dos resultados de cada ponto de dado $\mathbf{x}_i$ pertencente ao cluster C_j.
- $\mathbf{x}_i$: Um ponto de dado individual (vetor de variáveis).
- $\boldsymbol{\mu}_j$: O centroide (vetor de médias) do grupo j.
- $||\mathbf{x}_i - \boldsymbol{\mu}_j||^2$: Distância Euclidiana Quadrada entre o ponto $\mathbf{x}_i$ e o centroide $\boldsymbol{\mu}_j$.

O processo iterativo do algoritmo é iniciado com a escolha dos K clusters, sendo que cada centróide corresponde a um grupo. Os K centróides iniciais são atribuídos aleatoriamente. O algoritmo segue as etapas a seguir:

- **Inicialização:** Definição do K e posicionamento dos K centróides iniciais no espaço das variáveis.
- **Atribuição:** Cada observação se une ao grupo cujo centróide seja o mais próximo, fazendo uso da distância euclidiana quadrada.
- **Atualização:** Com os *clusters* formados, os centróides são recalculados, criando-se um novo ponto médio das observações que foram atribuídas ao respectivo cluster.

As iterações repetem-se até que a alocação das observações não se altere de forma significativa ou até que a posição dos centróides se estabilize, indicando a convergência do modelo.

2.3.2 Método de Clusterização Hierárquica

O método hierárquico adota uma formação de *clusters* distinta do K-Means. A diferença crucial reside no processo de clusterização: enquanto o K-Means é um método particional baseado em centróides, o agrupamento hierárquico é um método construtivo que estabelece uma estrutura em forma de dendrograma (árvore), a qual representa a relação de similaridade e distância entre as observações. Existem dois métodos principais segundo [7]: o Aglomerativo (*bottom-up*) e o Divisivo (*top-down*).

Para este trabalho, implementamos o método hierárquico aglomerativo, o mais comumente aplicado. Esta abordagem funciona de baixo para cima, onde cada observação é um grupo individual. Em seguida, ocorrem fusões iterativas mediante suas similaridades.

- **Inicialização:** Cada observação é tratada como um único grupo.
- **Fusão:** Em cada iteração, os *clusters* mais próximos, de maior similaridade e menor distância, são fundidos.
- **Conclusão:** O processo continua até que todas as observações estejam unidas em um grupo raiz.

Para a medição da proximidade entre os *clusters* a serem mesclados (critério de fusão), foi utilizado o método de ligação *Ward* (*Ward's linkage*) dentre os demais existentes. O método *Ward* busca a fusão que resulta no menor aumento da Soma dos Quadrados Intra-Cluster (*Within-Cluster Sum of Squares*, WCSS), ou seja, na menor perda de coesão interna [13].

Matematicamente, o método *Ward* é baseado no cálculo da distância Euclidiana Quadrada entre os centroides dos *clusters*. O aumento na WCSS ($\Delta WCSS$) ao fundir os *clusters* C_i e C_j é formalmente dado por (Eq. 4):

$$\Delta WCSS(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} ||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||^2 \tag{4}$$

onde:

- $\Delta WCSS(C_i, C_j)$: É o aumento na WCSS causado pela fusão dos *clusters* C_i e C_j.
- n_i, n_j: São, respectivamente, o número de observações nos *clusters* C_i e C_j.
- $\boldsymbol{\mu}_i, \boldsymbol{\mu}_j$: São os vetores centróides dos *clusters* C_i e C_j.
- $||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||^2$: É a Distância Euclidiana Quadrada entre os centróides.

O método Ward busca, em cada iteração, o par de *clusters* que resulta no menor valor de $\Delta WCSS$. Conforme a Equação 4, a expressão mostra que a WCSS é penalizada não apenas pela distância entre os centróides ($\boldsymbol{\mu}_i$ e $\boldsymbol{\mu}_j$), mas também pelo tamanho dos *clusters* (n_i e n_j), favorecendo a formação de *clusters* compactos e balanceados.

Este método é conhecido por sua tendência a gerar *clusters* com alta coesão interna, o que indiretamente auxilia na separação entre os *clusters*, sendo relevante para interpretação de perfis homogêneos. O critério de Ward é frequentemente escolhido por produzir *clusters* de tamanhos aproximadamente iguais, o que evita que *outliers* ou grandes *clusters* dominem a análise. Outro fator importante é a redução no tempo computacional para efetuar o cálculo [7].

2.4 Métricas de Avaliação

O agrupamento, por ser uma técnica de aprendizagem não supervisionada, demanda uma avaliação baseada em critérios que quantifiquem a qualidade da estrutura interna dos *clusters*, uma vez que não há um rótulo de verdade (*ground truth*). Desta forma, as métricas de validação interna têm o propósito de medir o quanto os *clusters* são coesos internamente e o quão bem estão separados entre si. As seguintes métricas foram utilizadas para a comparação de performance entre os modelos K-Means e Hierárquico e para a validação da escolha final do K.

2.4.1 Pontuação de Silhueta (*Silhouette Score*)

A Pontuação de Silhueta avalia o quão bem cada observação se ajusta ao seu próprio grupo comparada ao grupo mais próximo. O valor da silhueta para uma observação i é dado pela Equação 5 [14]. O valor da Pontuação de Silhueta varia de -1 a 1. Valores mais próximos de 1 são desejáveis, pois indicam que a observação está bem alocada e distante do grupo vizinho.

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{5}$$

onde:

- s_i: É a Pontuação de Silhueta para a observação i, com $i = 1, 2, \ldots, n$, sendo n o número total de observações.
- a_i: É a coesão da observação i. É definida como a distância média da observação $\mathbf{x}_i$ a todos os outros pontos no mesmo grupo (C_k). Quanto menor o valor de a_i, maior a coesão interna.
- b_i: É a separação da observação i. É definida como a menor distância média da observação $\mathbf{x}_i$ a todos os pontos de qualquer outro grupo (C_j), sendo C_j o grupo vizinho mais próximo.
- $\max(a_i, b_i)$: É o maior valor entre a_i e b_i, usado para normalizar o resultado.

2.4.2 Índice de Davies-Bouldin (*Davies-Bouldin Index, DBI*)

O Índice de Davies-Bouldin (DBI) quantifica a qualidade do agrupamento com base na razão entre a dispersão média intra-*clusters* e a distância entre os centróides dos *clusters*. O DBI é uma métrica de avaliação em que valores menores são desejáveis. Um valor baixo de DBI indica uma agrupamento superior, pois é resultado de *clusters* densos e com alta separação entre si.

O DBI é formalmente definido pela Equação 6 [15]:

$$DBI = \frac{1}{K} \sum_{i=1}^{K} \max_{i \neq j}(R_{ij}) \tag{6}$$

onde:

- K: É o número total de *clusters* (clusters).
- $\sum_{i=1}^{K}$: Símbolo de Somatório, que significa: some os resultados para cada grupo i, do primeiro ($i = 1$) até o último (K).
- $\max_{i \neq j}(R_{ij})$: A instrução para encontrar o valor máximo da similaridade R_{ij} para um grupo específico i, comparando-o com todos os outros *clusters* j.
- R_{ij}: A medida de similaridade entre o grupo i e o grupo j.

A similaridade entre dois *clusters*, R_{ij}, é a medida de similaridade entre o grupo i e o grupo j. Esta relação (razão de dispersão) é formalmente definida por:

$$R_{ij} = \frac{s_i + s_j}{d(\mathbf{c}_i, \mathbf{c}_j)} \tag{7}$$

onde:

- R_{ij}: É a medida de similaridade entre o grupo i e o grupo j.
- s_i e s_j: São, respectivamente, as medidas de dispersão interna (dispersão média) dos *clusters* i e j.
- $d(\mathbf{c}_i, \mathbf{c}_j)$: É a distância entre os centróides ($\mathbf{c}_i$ e $\mathbf{c}_j$) do grupo i e do grupo j.

2.4.3 Correlação Cofenética (*Cophenetic Correlation Coefficient, CCC*)

A Correlação Cofenética é uma métrica utilizada especificamente para avaliar a fidelidade da estrutura hierárquica construída (o dendrograma) em relação às distâncias originais entre os dados. O valor do CCC varia entre 0 e 1. Valores mais próximos de 1 são desejáveis, pois indicam que o agrupamento hierárquico preservou fielmente as distâncias originais entre os dados.

O *CCC* é calculado como a correlação de Pearson entre duas matrizes de distâncias, conforme a Equação 8:

$$CCC = \text{corr}(D, D_c) \tag{8}$$

onde:

- D: É a matriz de distâncias originais entre os N pares de observações do conjunto de dados.
- D_c: É a matriz de distâncias cofenéticas. A distância cofenética entre duas observações é definida como a altura do ramo no dendrograma no qual elas são unidas pela primeira vez.
- $\text{corr}(\cdot)$: Representa o Coeficiente de Correlação de Pearson.

3 Resultados e Discussão

3.1 Análise Exploratória de Dados

A Análise Exploratória de Dados constituiu uma etapa crucial para a compreensão do portfólio, o entendimento do comportamento das variáveis e a determinação das etapas de pré-processamento necessárias para otimizar a performance dos algoritmos de agrupamento.

A variável categórica (*tp_selagem*), após a codificação One-Hot Encoding foi devidamente preparada. Contudo, as variáveis numéricas demonstraram forte assimetria e multimodalidade. Essa característica exigiu a aplicação da transformação Box-Cox (Equação 1) como etapa preliminar a normalização dos dados (Equação 2), mitigando o impacto da escala no cálculo da distância euclidiana.

A multimodalidade evidenciada pelos múltiplos picos nas distribuições das variáveis numéricas reforça a hipótese de que o conjunto de dados é composto por grupos dimensionais distintos, o que valida a escolha por uma abordagem de agrupamento não supervisionado.

3.2 Avaliação e Definição do K

Visto que o agrupamento é uma técnica de aprendizagem não supervisionada, sem um rótulo de verdade (*ground truth*), a determinação do número ideal de grupos (K) é uma decisão estratégica balizada por métricas estatísticas e por premissas de negócio. A determinação do K para os métodos K-Means e Hierárquico baseou-se inicialmente em critérios estatísticos, cujos resultados evidenciaram a complexidade do portfólio:

•**Para o K-Means:** A análise da Inércia visualizada pelo Método do Cotovelo (*Elbow Method*) em função de diferentes valores de K (conforme Figura 4) sugeriu um ponto de inflexão na região de $K = 6$. Embora este valor seja estatisticamente eficiente para a redução da variância intra-grupo, ele foi considerado insuficiente para capturar a granularidade e a diversidade exigida pelo negócio.

•**Para o Agrupamento Hierárquico:** A aplicação da Regra de Mojena [16] para corte no dendrograma (Figura 5) sugeriu um número de grupos significativamente mais elevado, na ordem de $K = 47$. Este resultado indica a existência de alta granularidade e nichos dimensionais no portfólio, estando mais próximo das premissas do negócio.

3.3 Validação K ótimo

Essa disparidade entre as sugestões dos métodos (o $K = 6$ do Elbow versus o $K = 47$ da Regra de Mojena) expôs o dilema central do nosso trabalho. Para encontrar o equilíbrio ideal entre o rigor estatístico e a utilidade prática, recorremos às métricas de validação interna.

Para isso, foi utilizada a Pontuação de Silhueta (*Silhouette Score*, Seção 2.4.1), que mede formalmente a coesão e a separação dos agrupamentos. Para determinar o K que otimiza essa métrica, foi simulado um range de K que se estendeu de 2 até 100 grupos para ambos os métodos, K-Means e Hierárquico.

A Figura 6 representa a análise comparativa da Pontuação de Silhueta média em função do número de grupos (K). A análise detalhada dessa curva indicou os pontos de máxima qualidade estatística para cada método.

3.4 Cenários de Modelagem e Avaliação

Como conclusão da simulação aplicada e alinhado às premissas de negócio, foram estabelecidos os valores de K estatisticamente relevantes: $K = 47$, sugerido pela Regra de Mojena do método hierárquico, e $K = 65$, identificado como o ponto de máxima Pontuação de Silhueta para o método hierárquico. Embora ambos os valores tenham sido definidos por critérios de agrupamento hierárquico, eles serão aplicados ao método K-Means, e o desempenho será medido por métricas de validação interna.

Durante a Análise Exploratória de Dados, foi identificada uma forte similaridade entre as variáveis *diâmetro* e *comprimento*. Essa correlação entre variáveis motivou uma premissa crucial para a análise de robustez: a necessidade de avaliar o impacto da redundância de informação. Sendo assim, dois conjuntos de dados foram usados no agrupamento:

- **Conjunto de Dados 1:** Contém todas as variáveis, incluindo a variável *comprimento*.
- **Conjunto de Dados 2:** Exclui a variável *comprimento*.

Portanto, para validar a estabilidade dos modelos e a relevância das variáveis de entrada, foram executados 8 cenários de agrupamento. A inclusão de cenários sem a variável '*comprimento*' foi motivada pela necessidade de avaliar o impacto desta variável na qualidade do agrupamento, dado seu potencial como fonte de alta variabilidade e sua similaridade com a variável *diâmetro*.

Os cenários de agrupamento foram definidos conforme a Tabela 2:

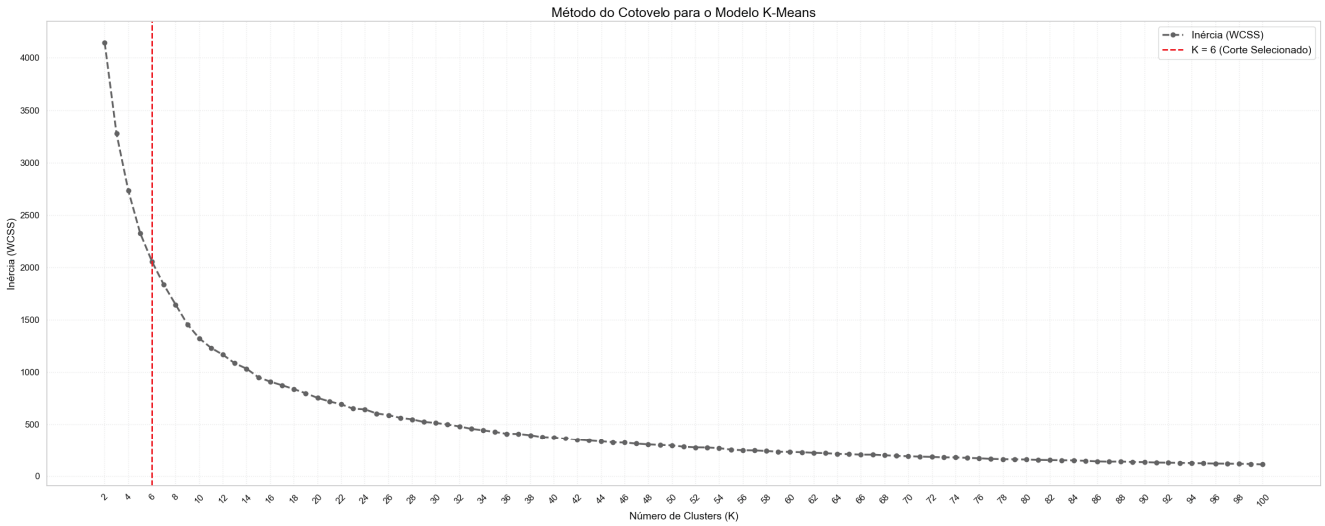

Figura 4: Determinação do número ótimo de grupos (K) através do Método Elbow. Fonte: O autor (2025)

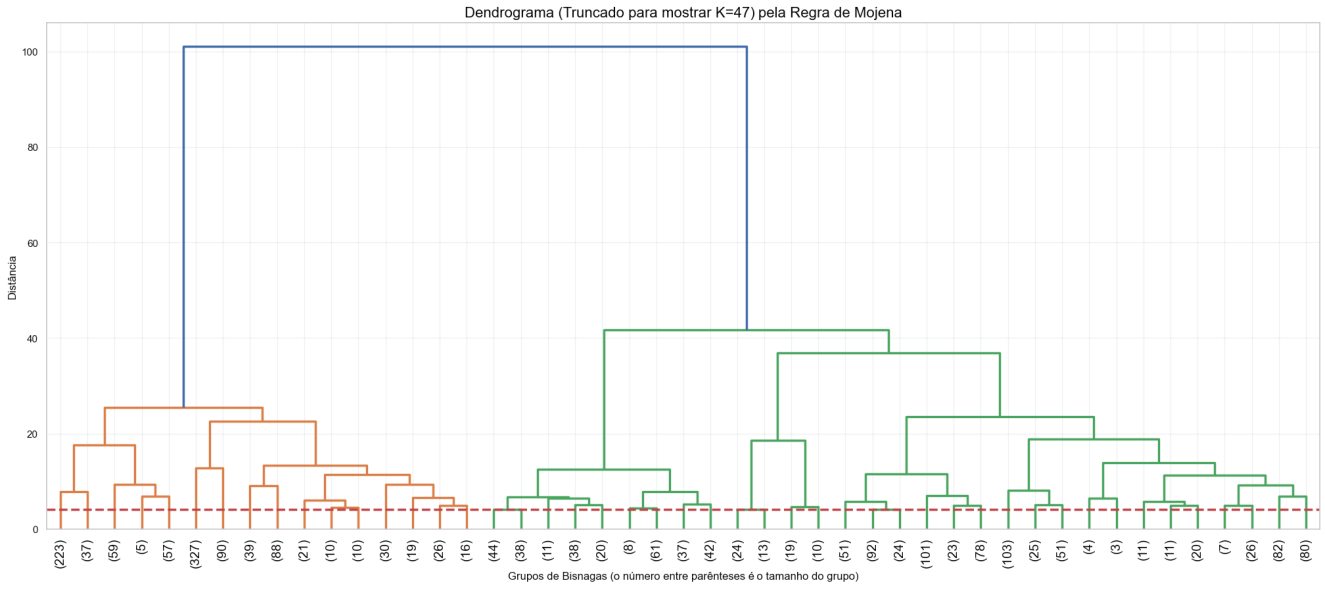

Figura 5: Determinação do número ótimo de grupos (K) através da Regra de Mojena aplicada ao Dendrograma. Fonte: O autor (2025)

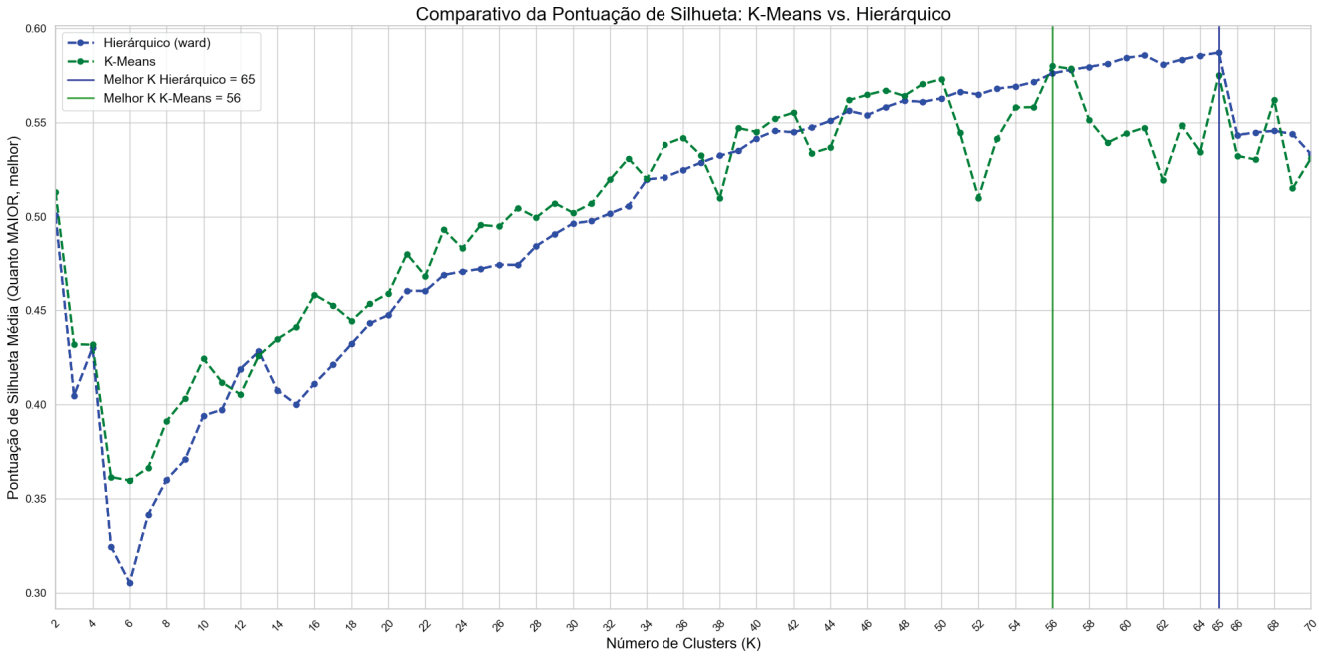

Figura 6: Análise comparativa da Pontuação de Silhueta média em função do número de grupos (*K*). Fonte: O autor (2025)

Tabela 2: Cenários de agrupamento e conjuntos de dados a serem utilizados.

Método	**K (grupos)**	**Conj. de Dados**	**Total**
K-Means	47	Conjunto 1 Conjunto 2	2
	65	Conjunto 1 Conjunto 2	2
Hierárquico	47	Conjunto 1 Conjunto 2	2
	65	Conjunto 1 Conjunto 2	2

A (Tabela 3) apresenta o desempenho dos 8 cenários de experimento em relação as métricas de validação interna: Pontuação de Silhueta (Silhouette Score, que deve ser maximizada), Índice de Davies-Bouldin (DBI, que deve ser minimizado) e Correlação Cofenética (aplicável apenas ao método Hierárquicos).

Ao observarmos os grupos gerados, torna-se possível identificar o ganho mais relevante de performance nas métricas de dispersão e separação quando agrupados com o Conjunto de Dados 2 (sem a variável '*comprimento*').

O Índice de Davies-Bouldin (*DBI*), que mede a razão entre a dispersão intra-grupos e a separação inter-grupos, diminuiu consideravelmente em todos os cenários com a exclusão da variável '*comprimento*'. Essa melhora validou a premissa de redundância, indicando que a remoção dessa variável reduziu a variância interna dos grupos.

No entanto, a Pontuação de Silhueta (*Silhouette Score*), que avalia a coesão e a separação em conjunto, apresentou um resultado misto:

- **Modelos Hierárquicos (C2 vs. C1 e C4 vs. C3):** O *Silhouette Score* teve uma queda, indicando que a eliminação da variável, embora melhorando o *DBI* (redução da dispersão), reduziu ligeiramente a coesão do agrupamento.
- **Modelos K-Means (C6 vs. C5 e C8 vs. C7):** O *Silhouette Score* apresentou uma melhora, ainda que marginal, sugerindo que a remoção da variável redundante não prejudicou, e até auxiliou, a coesão do agrupamento.

Tabela 3: Desempenho das métricas de validação interna para os 8 cenários de agrupamento.

Cen.	Método	K	Conj. Dados	Silhueta	DBI	Corr. Cofenética
C1	Hierárquico	47	Conjunto 1	0,5582	0,8128	0,7216
C2	Hierárquico	47	Conjunto 2	0,5578	0,6119	0,8060
C3	Hierárquico	65	Conjunto 1	0,5872	0,7268	0,7216
C4	Hierárquico	65	Conjunto 2	0,5746	0,5195	0,8060
C5	K-Means	47	Conjunto 1	0,5386	0,7893	N/A
C6	K-Means	47	Conjunto 2	0,5453	0,5688	N/A
C7	K-Means	65	Conjunto 1	0,5728	0,6218	N/A
C8	K-Means	65	Conjunto 2	0,5733	0,5240	N/A

3.5 Performance e Escolha Final

A avaliação de desempenho dos cenários propostos, sumarizados na Tabela 3 , demandou uma análise que considerasse não apenas as métricas estatísticas, mas também as premissas importantes de negócio e a viabilidade operacional.

O Cenário C3 (Clusterização Hierárquica com $K = 65$, Conjunto Completo) alcançou o maior valor absoluto de Pontuação de Silhueta (0,5872), indicando maior coesão interna. Contudo, essa solução apresentou um Índice de Davies-Bouldin (DBI) de 0,7268, o mais alto entre os cenários com K=65, o que sugere uma dispersão excessiva entre os clusters.

Em contraste, o Cenário C8 (K-Means com $K = 65$, sem a variável comprimento) foi selecionado como o modelo final. Esta escolha foi fundamentada em um balanceamento estratégico: o C8 registrou um elevado Silhouette Score (0,5733) e, crucialmente, um DBI baixo (0,5240), indicando a melhor separação e menor dispersão entre os grupos.

Adicionalmente, o C8 demonstrou a melhor coerência na validação técnica dos perfis para o processo produtivo final, superando os modelos hierárquicos em viabilidade operacional. O algoritmo K-Means com a remoção da variável redundante provou ser a solução mais robusta para manter a coesão e reduzir drasticamente a dispersão, alinhando a performance estatística às necessidades de negócio.

3.6 Interpretação do método empregado

O modelo final selecionado (K-Means com K=65, Cenário C8) obteve 65 grupos de bisnagas. O primeiro passo na interpretação dessas famílias é a análise da distribuição (Figura 7) do volume total de SKUs, crucial para a estratégia de sequenciamento produtivo. Como pode-se notar, a distribuição dos grupos não obteve uma homogeneidade, isso já era esperado dado o desbalanceamento da variável categórica (tp_selagem). Chama atenção também os grupos com um número baixo de observações (Sku's), que expressam a representação de Skus bem peculiares.

4 Conclusão

O presente estudo aplicou técnicas de aprendizagem de máquina (clusterização) com o objetivo de solucionar o problema da segmentação subjetiva de SKUs de bisnagas em uma indústria de cosméticos. A abordagem demonstrou-se eficaz ao eliminar a dependência da experiência subjetiva, conferindo maior celeridade e padronização ao processo decisório. O objetivo principal foi plenamente atingido com a entrega do modelo K-Means, com $K = 65$ grupos, que define 65 famílias homogêneas e estatisticamente robustas de bisnagas, traduzindo as características dimensionais e de processo em informações acionáveis.

O principal desafio e objeto de contribuição metodológica foi a equalização e a definição do K ótimo. O rigor empregado na análise de robustez dos 8 cenários (Tabela 3) permitiu uma análise ponderada entre a otimização estatística e a viabilidade operacional, resultando na escolha do modelo que oferece o melhor balanceamento estratégico. Esse agrupamento em 65 famílias fornece subsídios diretos para a gestão industrial, permitindo, através da comparação entre *clusters* e *SKUs*, a minimização de *setups* e a otimização do sequenciamento produtivo, entre outras ações. Tais iniciativas impactam positivamente o OEE (*Overall Equipment Effectiveness*) das linhas de envase.

Como limitação do estudo, reconhece-se que não foram explorados outros métodos de agrupamento avançados, como DBSCAN ou *Gaussian Mixture Models* (GMMs), cuja aplicação poderia refinar a segmentação dos grupos de nicho. Para trabalhos futuros, sugere-se a expansão da análise metodológica com o teste de novos modelos (como DBSCAN ou GMMs) para refinar a estrutura de agrupamento.

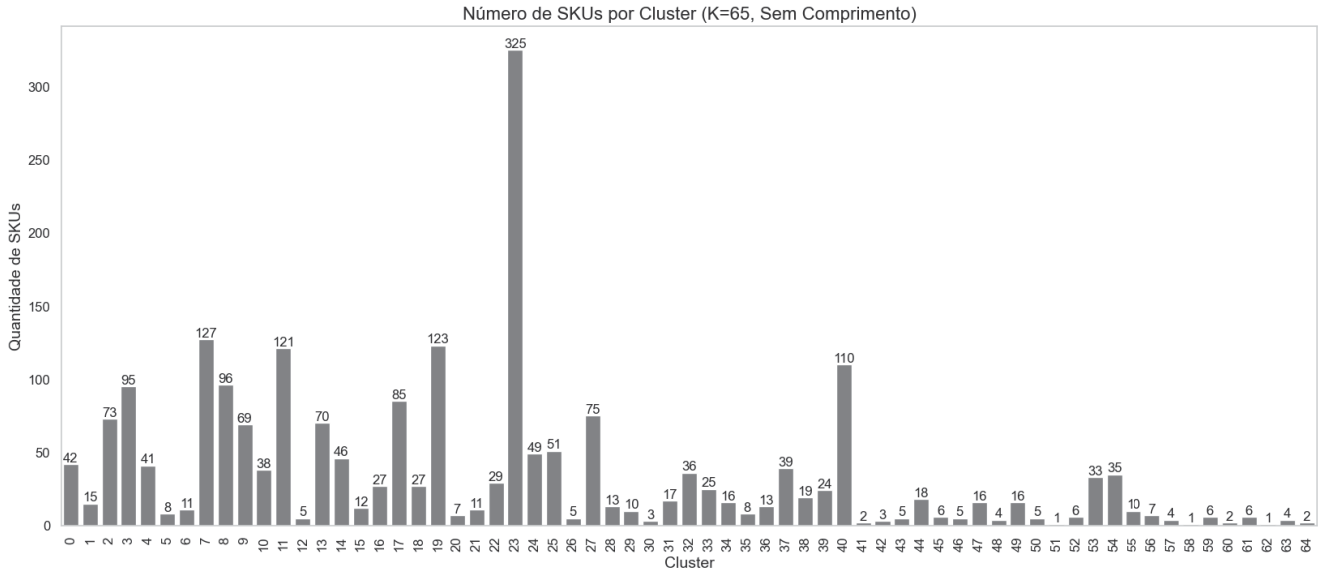

Figura 7: A Figura 5 Demonstra a distribuição do volume de SKUs (número de observações) em cada um dos 65 grupos gerados. Fonte: O autor (2025)

Agradecimentos

Em primeiro lugar, dedico a Deus minha profunda gratidão pela sabedoria e força concedidas ao longo desta jornada acadêmica.

Meu agradecimento especial e inestimável é direcionado à minha família, pilares essenciais desta conquista. À minha esposa, Fernanda Renner, e à minha filha, Catarina Renner, por serem a inspiração diária e a fonte inesgotável de motivação que tornaram possível a dedicação integral à conclusão deste projeto.

Estendo meus sinceros agradecimentos à Universidade Federal do Paraná (UFPR) e, de forma particular, ao corpo docente do Programa de Pós-Graduação em Ciência de Dados e Big Data. O conhecimento técnico de excelência e o rigor metodológico providos foram elementos fundamentais que garantiram a qualidade e o desenvolvimento científico deste trabalho.

Referências

[1] ASSOCIAÇÃO BRASILEIRA DE SUPERMERCADOS (ABRAS). Setor de beleza e cuidados pessoais deve crescer 7% ao ano até 2027 no brasil, segundo estudo, June 2024.

[2] M. S. Santos, A. G. L. Suela, A. O. S. Góes, M. A. L. Costa, S. S. dos Reis, and S. S. de Jesus. A estratégia competitiva de inovação na indústria 5.0: ideias, provocações e reflexões. *Revista de Gestão e Secretariado*, 16(1):e4586, 2025.

[3] ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA DE HIGIENE PESSOAL, PERFUMARIA E COSMÉTICOS (ABIHPEC). Panorama global de consumo de produtos de higiene pessoal, perfumaria e cosméticos marcam o primeiro dia da semana abihpec de mercado 2022, 2022.

[4] Roniel Venâncio Santana and Heráclito Lopes Jaguaribe Pontes. Aplicação da clusterização por k-means para criação de sistema de recomendação de produtos baseado em perfis de compra. *Navus - Revista de Gestão e Tecnologia*, 10:01–14, 2020.

[5] H. Allende-Cid. Machine learning: catalisador da ciência. *Computação Brasil*, (39):15–18, 2019.

[6] Cledjan Torres da Costa and outros. União de dados por clusterização para construção de modelos de predição de evasão. *RENOTE*, 21(2):393–402, 2023.

[7] Jean Metz. Interpretação de clusters gerados por algoritmos de clustering hierárquico. Dissertação (mestrado em ciências de computação e matemática computacional), Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, São Carlos, 2006.

[8] IFSULDEMINAS, editor. *Aplicação do algoritmo K-means no agrupamento de imagens de satélite para identificação de seleções em clusters*, volume 15, Pouso Alegre, MG, 2023. IFSULDEMINAS.

[9] Wagner Vidal Xavier da Silva. O impacto do uso da biblioteca pandas do python como ferramenta de análise de dados referente aos casos graves de covid-19 notificados pela secretaria estadual de saúde do recife entre julho de 2021 a junho de 2022. In *Anais do VI Simpósio de Inovação em Engenharia Biomédica - SABIO 2022*, page 22, Recife, PE, 2022. Editora Universitária da UFPE.

[10] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

[11] Romário Parreira Pita and outros. ComparaÇÃo entre encoders para precificaÇÃo de imÓveis por meio do uso de aprendizado de mÁquina. In *Anais do 27º Encontro Nacional de Modelagem Computacional (ENMC) e 15º Encontro de Ciência e Tecnologia de Materiais (ECTM)*, Ilhéus, BA, 2024. Even3.

[12] Edric Brasileiro Troccoli and outros. K-means clustering using principal component analysis to automate label organization in multi-attribute seismic facies analysis. *Journal of Applied Geophysics*, 198:104555, 2022.

[13] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[14] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987.

[15] David L. Davies and Don L. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1:224–227, 1979.

[16] Richard Mojena. Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal*, 20(4):359–363, 1977.