Lista de Exercícios 2 – Análise de Dados IV
Marcos Rodrigues de Oliveira Júnior

1) É uma técnica de aprendizado não supervisionado que agrupa um conjunto de objetos em clusters (grupos) com base em suas similaridades.

2) O principal objetivo é desenvolver uma taxonomia que particione objetos em grupos com percepções similares

Ou seja, é identificar padrões e estruturas nos dados, agrupando elementos similares em conjuntos (conglomerados) distintos. Essa técnica é utilizada para entender melhor os dados, reduzir complexidade e tomar decisões estratégicas com base nos grupos identificados.

3) Podemos usar na identificação de grupos de investimentos de acordo com perfis de ricos. Identificar segmentos homogêneos de consumidores e estabelecer programas de marketing específicos para cada segmento. Ou identificar grupos de alunos mais propensos à evasão escolar por exemplo.

4) Distância Euclidiana: A distância entre duas observações (i e j) corresponde à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações (i e j) para todas as p variáveis:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

Em que xik é o valor da variável k referente à observação i e xjk para a observação j. Nesta abordagem, quanto menor a distância, mais similares serão as observações.

Distância Quadrática Euclidiana: a distância entre duas observações (i e j) corresponde à soma dos quadrados da diferença entre i e j para todas as p variáveis (Recomendada para os métodos de agrupamento centroide e Ward):

$$d_{ij}{}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

Distância de Minkowski:

$$d_{ij} = \left(\sum_{k=1}^{p} (|x_{ik} - x_{jk}|)^n\right)^{1/n}$$

Em que dij é a distância de Minkowski entre as observações i e j, p é o número de variáveis, e n = 1, 2, ..., ∞.

Distância de Mahalanobis:

$$d_{ij} = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)}$$

Em que S é a estimativa amostral da matriz de variância covariância Σ dentro dos agrupamentos

5) Método Hierárquico Aglomerativo (Bottom-up)

Esse método começa considerando cada ponto como um cluster isolado e, aos poucos, vai mesclando os mais próximos até formar um único cluster contendo todos os elementos.

Principais métodos:

Vizinhos Mais Próximos (Single Linkage)

Aqui a distância entre dois clusters é definida pela menor distância entre qualquer par de pontos. Pode formar clusters alongados e encadeados e é sensível a outliers.

Vizinhos Mais Distantes (Complete Linkage)

Usa a maior distância entre qualquer par de pontos de dois clusters, produz cluster mais compactos e esféricos e é menos sensível a outliers.

Média das Distâncias (Average Linkage)

Calcula a média de todas as distâncias entre os pontos de dois clusters equilibrando características do Single e do complete linkage.

Centroide dos Clusters (Centroid Linkage)

Este método define a coordenada de cada grupo como sendo a média das coordenadas de seus objetos. Uma vez obtida essa coordenada, denominada centroide, a distância entre os grupos é obtida através do cálculo das distâncias entre os centroides. Mede a distância entre os centroides (médias dos pontos) de dois clusters.

Método Ward's

O método de Ward busca unir objetos que tornem os agrupamentos formados os mais homogêneos possível. A medida de homogeneidade utilizada baseia-se na partição da soma de quadrados total de uma análise de variância.

O método Ward funciona minimizando a soma dos quadrados das distâncias dentro dos clusters (similar à lógica do k-means). Isso quer dizer que:

- Ele puxa os agrupamentos para tentar manter a variância interna pequena.
- Se houver um outlier muito distante, ele aumenta muito a variância, e o algoritmo tenta isolá-lo rapidamente, o que pode distortar o agrupamento geral.

6) **Comparação entre Métodos Hierárquicos e Não Hierárquicos na Análise de Clusters**

Critério	**Método Hierárquico**	**Método Não Hierárquico**
Definição prévia do número de clusters	Não precisa definir k previamente	Precisa definir k antes de rodar o algoritmo
Forma dos clusters	Pode gerar clusters de diferentes formas e tamanhos	Tende a formar clusters esféricos (ex: K-Means)
Interpretação	Produz um dendrograma, facilitando a análise dos agrupamentos	Não gera estrutura hierárquica, tornando a interpretação mais difícil

Flexibilidade	Permite fusão e divisão gradual de clusters	Depende mais de parâmetros iniciais (ex: k no K-Means)
Escalabilidade	Computacionalmente pesado para grandes conjuntos de dados ($O(n^2)$ ou pior)	Mais eficiente em grandes bases de dados O(n) para alguns métodos como K-Means)
Sensibilidade a Outliers	Pode ser afetado por outliers (dependendo do método)	Alguns métodos, como DBSCAN, lidam bem com ruídos e outliers
Aplicação em dados grandes	Desempenho ruim em grandes volumes de dados	Mais indicado para grandes bases de dados

7) É interessante usar os métodos hierárquicos quando (bases pequenas e exploratórias):

- O número de clusters não é conhecido e você quer explorá-los visualmente.
- Precisa de interpretação clara por meio de um dendrograma.
- A base de dados não é muito grande (cerca de milhares de pontos, no máximo).
- Deseja um método mais robusto para diferentes formas de clusters.

E usar os métodos não hierárquicos quando (grandes volumes de dados e melhor eficiência):

- Está lidando com grandes conjuntos de dados, pois esses métodos são mais eficientes.
- Já tem uma ideia do número de clusters esperados (ex: segmentação de clientes).
- Os clusters são aproximadamente esféricos (como K-Means).
- Quer um método mais rápido e escalável, como DBSCAN para detectar outliers.

8) Sim, os métodos hierárquicos e não hierárquicos podem ser considerados complementares, pois cada um tem vantagens e desvantagens que podem ser combinadas para melhorar a análise de clusters. Podemos usar a exploração inicial com o método hierárquico já que seus métodos ajudam a entender a estrutura dos dados e podem ser usados para determinar o número ideal de cluster por meio da

análise de dendrograma. Após isso usar um método não hierárquico com o K-means por exemplo, para segmentar os dados de forma mais eficiente e escalável. Assim o uso combinado permite minimizar desvantagens individuais. Métodos hierárquicos criam agrupamentos mais estruturados enquanto os não hierárquicos são mais fáceis e rápidos para ajustar os clusters conforme novas informações são adicionadas.

9) A determinação do número ideal de clusters em uma análise de conglomerados pode ser feita por diferentes técnicas, dependendo do método utilizado. Em métodos hierárquicos, como o de Ward, a análise do dendrograma é uma abordagem comum. O pesquisador observa a altura das fusões entre clusters e define um corte onde há uma grande variação, indicando a separação natural dos grupos.

Nos métodos não hierárquicos, como o K-Means, uma das técnicas mais utilizadas é o Método do Cotovelo (Elbow Method). Ele consiste em calcular a soma das distâncias quadradas intra-cluster (Inertia/WCSS – Within-Cluster Sum of Squares) para diferentes valores de kkk. Ao plotar um gráfico variando kkk, o ponto onde a curva forma um "cotovelo" representa o número ideal de clusters, pois a partir desse ponto a redução da variabilidade dentro dos clusters se torna menos significativa.

Outra métrica importante é o Índice de Silhueta (Silhouette Score), que avalia a qualidade do agrupamento ao medir o quão próximo um ponto está do seu próprio cluster em comparação com os outros clusters. Esse índice varia de -1 a 1, onde valores próximos de 1 indicam clusters bem definidos, enquanto valores próximos de 0 sugerem que os clusters estão sobrepostos. O número ótimo de clusters será aquele que maximiza esse índice.

O Gap Statistic é outro método que pode ser utilizado para definir kkk, comparando a dispersão dos clusters obtidos com a dispersão esperada de um conjunto de dados aleatórios. O número ideal de clusters é aquele que maximiza esse critério.

Além dessas abordagens, existem critérios estatísticos como o Bayesian Information Criterion (BIC) e o Akaike Information Criterion (AIC), que são frequentemente usados em modelos probabilísticos, como o Gaussian Mixture Models (GMM). Esses critérios avaliam a qualidade do modelo de clusterização, onde o menor valor de BIC ou AIC indica o número ótimo de clusters.

A escolha do método mais adequado depende do tipo de dados e do modelo de clusterização adotado. Métodos hierárquicos se beneficiam da análise do dendrograma, enquanto o K-Means costuma ser combinado com o Método do Cotovelo e o Índice de Silhueta. Modelos baseados em densidade, como o

DBSCAN, utilizam parâmetros diferentes, como a densidade e a distância mínima entre pontos, para definir os grupos de forma mais flexível.

10) A padronização das variáveis é essencial em muitas situações antes da aplicação da análise de conglomerados, pois garante que todas as variáveis tenham a mesma influência na formação dos clusters. Isso ocorre porque muitos algoritmos de agrupamento utilizam medidas de distância, como a distância Euclidiana, que são sensíveis à escala das variáveis.

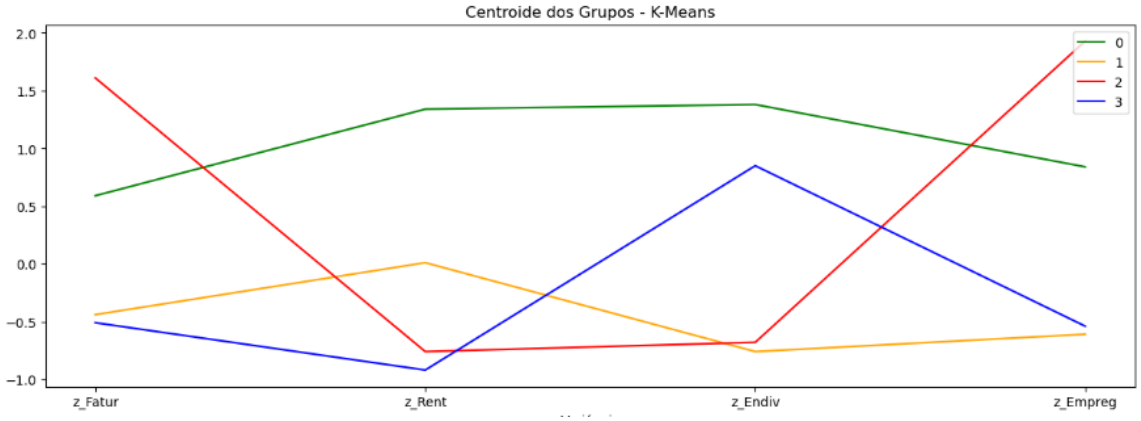
A padronização é necessária quando:

- **As variáveis estão em escalas diferentes**: Se os dados possuem unidades diferentes (por exemplo, altura em centímetros e peso em quilogramas), variáveis com valores maiores podem dominar a formação dos clusters.
- **Os dados possuem distribuições muito diferentes**: Algumas variáveis podem ter variações muito maiores do que outras, distorcendo o cálculo das distâncias entre os pontos.
- **Métodos baseados em distância são usados**: Algoritmos como K-Means, Hierárquico (com distância Euclidiana) e DBSCAN são altamente influenciados por variáveis de grande escala.

Assim, todas as variáveis contribuem igualmente para a clusterização, permitindo uma segmentação mais equilibrada e representativa dos dados.

11) Analisando com todas as empresas, obtivemos pela analise hierárquica que 3 ou 4 clusters seriam uma opção para a produção da análise pelo k-means, escolhi 4 resultando em:

	c1	c2	dist	n	heterogeneidade
14	28.0	36.0	3.317625	8.0	1.263365
15	10.0	35.0	3.359539	3.0	28.285784
16	1.0	2.0	4.309811	2.0	31.629855
17	34.0	38.0	5.672998	6.0	28.489272
18	32.0	39.0	7.289194	11.0	10.678882
19	0.0	37.0	8.067598	3.0	14.440019
20	40.0	41.0	9.232561	5.0	69.764195
21	42.0	43.0	15.673583	17.0	44.283520
22	44.0	46.0	22.614397	20.0	29.832745
23	45.0	47.0	29.360893	25.0	NaN



Onde temos uma boa heterogeneidade entre os grupos e homogeneidade dentro deles, sendo:

Cluster 1: Alta performance geral – empresas com alto faturamento, rentabilidade, endividamento moderado-alto e muitos empregados. Provavelmente são empresas grandes e maduras.

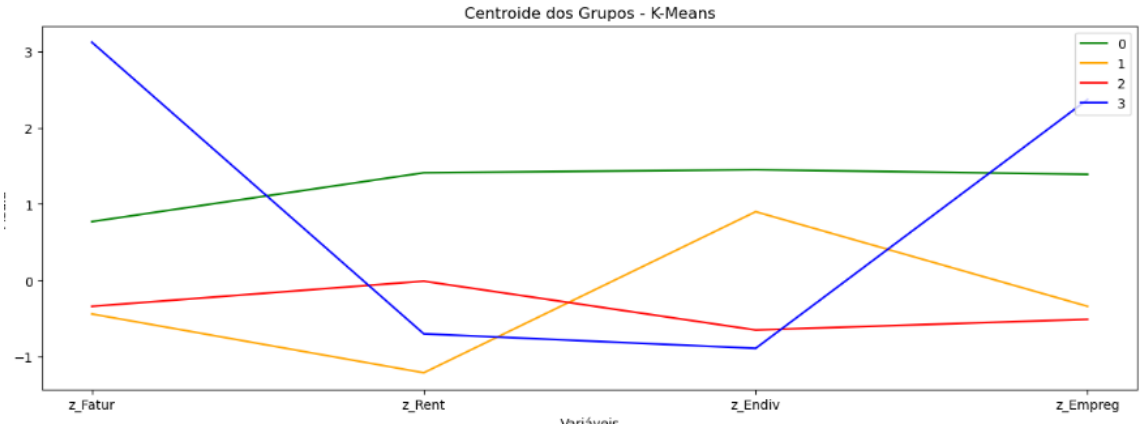
Cluster 2: Empresas com alta estrutura, mas pouco retorno – muito faturamento e muitos empregados, mas baixa rentabilidade e pouco endividamento. Pode indicar ineficiência operacional ou empresas com potencial pouco explorado.

Cluster 3: Empresas médias ou discretas – nada se destaca muito. Talvez empresas mais equilibradas, sem grandes riscos nem grandes destaques. Uma espécie de "grupo neutro" ou estável.

Cluster 4: Empresas muito endividadas, com baixo faturamento, baixa rentabilidade e poucos empregados. Provável grupo de alto risco financeiro.

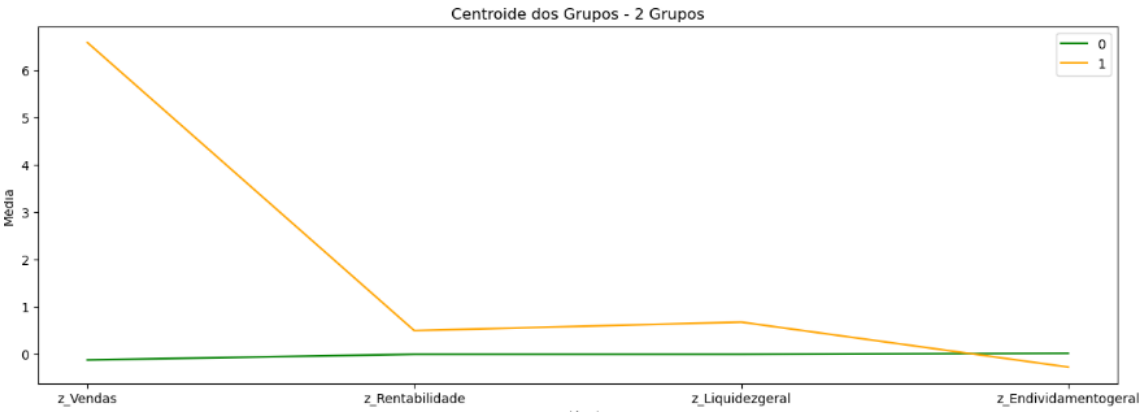
Agora, ao tirar as 3 maiores empresas (Empresas que tem maior número de empregados e Faturamento), retiramos as empresas CBA, V % M Do Brasil e Alcoa, temos (continuando com 5 clusters)

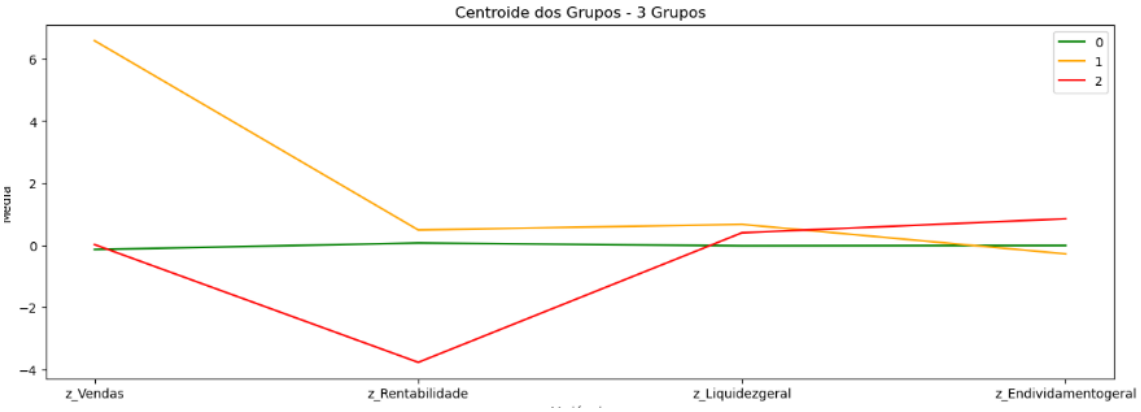
	c1	c2	dist	n	heterogeneidade
11	15.0	30.0	2.164895	4.0	13.380902
12	27.0	28.0	2.454577	6.0	79.616992
13	4.0	31.0	4.408838	4.0	3.812223
14	1.0	2.0	4.576913	2.0	14.401913
15	33.0	34.0	5.236076	10.0	43.851656
16	29.0	37.0	7.532182	13.0	31.138224
17	32.0	36.0	9.877569	4.0	54.493109
18	35.0	38.0	15.260164	17.0	62.469559
19	39.0	40.0	24.793121	21.0	16.813069
20	0.0	41.0	28.961606	22.0	NaN



De modo geral, a ausência das três maiores empresas fez com que os extremos fossem suavizados, permitindo uma leitura mais clara dos padrões entre as empresas médias. Além disso, o perfil dos clusters ficou mais equilibrado e permitiu identificar diferentes combinações entre estrutura, rentabilidade e endividamento — revelando, por exemplo, grupos com bom desempenho geral, outros com estrutura, mas pouca eficiência, e alguns em situação mais frágil.

12) Pelo método hierárquico observamos que 2 ou 3 grupos são possíveis boas quantidades, obtendo:





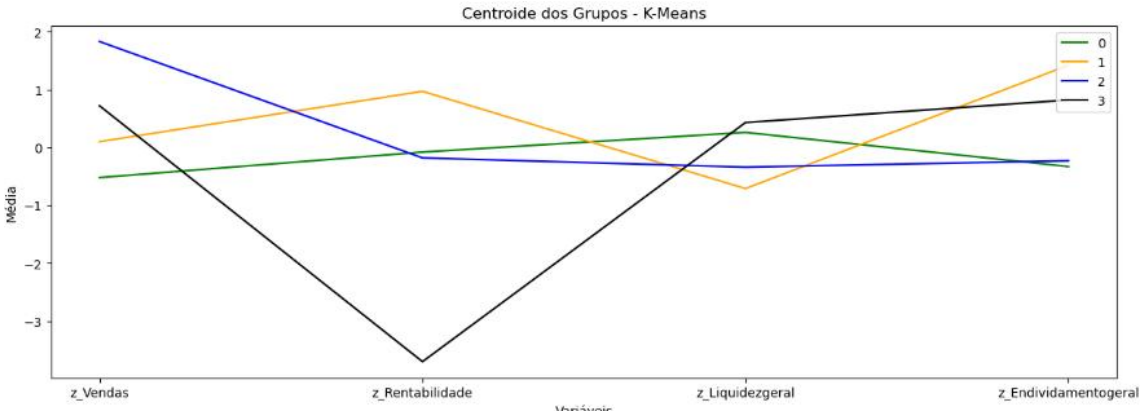
E escolhendo fazer o método Kmeans de não hierárquico com 3 grupos resultando numa tabela anova

Variável	QMC	df1	QME	df2	F	p-valor
z_Vendas	22.107	2	0.123	47	179.599	0.000
z_Rentabilidade	7.344	2	0.751	47	9.774	0.000
z_Liquidezgeral	0.314	2	1.050	47	0.299	0.743
z_Endividamentogeral	0.401	2	1.047	47	0.383	0.684

Que nos mostra que a variável mais determinando é a de Vendas, e que pelas proporções ela está puxando toda a análse, visto que havendo os 3 grupos, há apenas 1 no grupo 2 e 1 no grupo 3, ou seja, outliers.

13) Agora, retirando as 2 maiores empresas (Petrobras e Distribuidora Pretrobras) conseguimos fazer uma análise melhor (seria interessante também excluir a fiat)

A partir disso então pelo método hierárquico uma boa escolha é 5 ou 4 grupos, faremos com 4, obtendo então:



E uma tabela anova:

Variável	QMC	df1	QME	df2	F	p-valor
z_Vendas	11.962	3	0.275	44	43.442	0.000
z_Rentabilidade	7.253	3	0.596	44	12.163	0.000
z_Liquidezgeral	2.372	3	0.929	44	2.552	0.068
z_Endividamentogeral	6.847	3	0.624	44	10.972	0.000

Aonde ainda assim, as vendas importam muito.

14) Após a construção do dendrograma, foi possível identificar visualmente a formação de agrupamentos com base na distância entre os alunos. Observa-se que existe uma separação clara entre dois grandes grupos principais: o primeiro, formado pelos alunos A, B, C e F, reúne aqueles com maiores níveis de satisfação geral com o curso, apresentando notas elevadas nos três quesitos avaliados. O segundo grupo reúne os demais alunos, com destaque para D, G, H e J0, que apresentaram escores mais baixos, indicando um menor grau de satisfação.

Além desses dois grupos principais, também é possível observar a existência de um subgrupo intermediário, composto por alunos como E e I, que demonstram satisfação moderada. Essa análise permite concluir que os alunos podem ser agrupados em dois ou três conglomerados, conforme o critério adotado para o corte do dendrograma. A segmentação dos alunos com base em seu nível de satisfação é útil para identificar perfis distintos e orientar melhorias no curso de forma mais direcionada às necessidades percebidas pelos estudantes.

15) A análise de conglomerados foi aplicada a um conjunto de dados composto por 25 empresas varejistas, levando em conta três variáveis principais: número de itens no sortimento, quantidade de lojas com mais de 1.000 m^2 e faturamento mensal em reais. Antes de aplicar os métodos de clusterização, todas as variáveis foram padronizadas utilizando o método de Z-scores, o que garante que cada variável

tenha média zero e desvio padrão igual a um, evitando que variáveis com diferentes escalas influenciem desproporcionalmente os resultados.

Inicialmente, foi realizada a análise de conglomerados hierárquicos utilizando a distância quadrática euclidiana como métrica de dissimilaridade e o método Between Groups como critério de ligação. O dendrograma resultante revelou uma estrutura de agrupamento clara entre as empresas, com a formação de três a quatro conglomerados distintos, dependendo do ponto de corte considerado. Observa-se, por exemplo, a presença de um grupo formado por empresas com porte muito elevado, como as empresas S, U e Y, que se destacam com altíssimo sortimento de itens, maior número de lojas de grande porte e faturamento significativamente superior à média. Estas empresas compõem um grupo de grandes redes varejistas.

Um segundo conglomerado agrupa empresas de porte intermediário, como F, G, K, P, Q e V, que apresentam um sortimento amplo e bons níveis de faturamento e infraestrutura, porém em menor escala que o primeiro grupo. Já um terceiro grupo pode ser composto por empresas menores, como M, N, O e A, com baixo número de itens, poucas lojas e faturamento reduzido.

Para complementar a análise, foi aplicado o método de K-means clustering com a escolha de três clusters, alinhando-se à estrutura sugerida pelo dendrograma. A segmentação reforçou os agrupamentos obtidos pela análise hierárquica, alocando as empresas de maneira coerente com seus perfis de atuação. O cluster 1 agrupou empresas de pequeno porte, o cluster 2 empresas intermediárias e o cluster 3 as grandes redes. A análise dos centróides revelou diferenças marcantes entre os grupos, principalmente no faturamento e na quantidade de itens ofertados.

Em síntese, a análise de conglomerados permitiu identificar padrões distintos entre as empresas varejistas da amostra, facilitando a segmentação do mercado em três grandes perfis: pequenas, médias e grandes empresas. Essa segmentação pode ser valiosa para fins de benchmarking, definição de estratégias comerciais e compreensão da estrutura do setor.

16)

a) Inicialmente, aplicou-se a técnica de análise hierárquica de conglomerados com a distância quadrática euclidiana e o método do vizinho mais próximo (Nearest Neighbor ou Single Linkage). Esse método tende a formar aglomerados alongados, pois liga os grupos a partir das observações mais próximas entre si. O dendrograma gerado permitiu a visualização das junções passo a passo, e a análise visual sugeriu um número apropriado de conglomerados. Complementando a análise, aplicou-se o método K-means com base nas observações anteriores para confirmar a

estrutura dos grupos. A comparação das saídas mostrou consistência, especialmente ao observar clusters formados por empresas com alto faturamento, grande área e número expressivo de funcionários. A recomendação, com base nas junções e nos perfis semelhantes, seria a formação de três conglomerados principais: grandes redes nacionais, empresas de médio porte regionais e supermercados locais com estrutura mais enxuta.

b) A análise foi refeita utilizando o método de ligação entre grupos (Between Groups ou média intergrupos) com a mesma métrica de distância. Esse método tende a gerar conglomerados mais balanceados, minimizando a variabilidade interna dos clusters. A estrutura de agrupamento se manteve coerente com a anterior, porém observou-se uma organização mais estável entre os clusters intermediários. O dendrograma mostrou uma melhor separação entre conglomerados de médio porte, sugerindo novamente três conglomerados com forte distinção entre grupos de alta, média e baixa estrutura empresarial.

c) Em uma terceira abordagem, aplicou-se o método do vizinho mais distante (Furthest Neighbor ou complete linkage), utilizando a correlação de Pearson como medida de similaridade. Antes, os dados foram padronizados por média 1. Esse método tende a formar grupos mais compactos e com maior separação entre si. A métrica de correlação de Pearson mede a associação linear entre as variáveis, sendo mais robusta a escalas. O resultado foi a formação de conglomerados com empresas mais correlacionadas em comportamento geral, mesmo que com valores absolutos diferentes. Empresas com crescimento proporcional em todas as variáveis foram agrupadas, mesmo com valores diferentes em magnitude.

d) A análise foi repetida, agora com a exclusão das três maiores empresas (Carrefour, Pão de Açúcar e Walmart), que eram outliers em termos de faturamento e estrutura. A exclusão permitiu observar a estrutura interna dos grupos restantes com maior clareza. A partir do gráfico Icicle, foi possível visualizar o comportamento de fusão dos grupos ao longo do processo hierárquico. Ao considerar quatro conglomerados como solução, observou-se uma separação mais equilibrada entre os grupos. As empresas se dividiram em clusters com perfil regional, grandes redes nacionais em ascensão e redes locais.

e) Analisando o esquema de aglomeração até o estágio 12 (inclusive), foi possível descrever passo a passo como as empresas foram sendo agrupadas. Inicialmente, pares de empresas muito semelhantes (com distância mínima) se uniram, como no caso de supermercados com porte semelhante em área e número de funcionários. A cada estágio, os agrupamentos iam incorporando novas empresas ou se fundindo com outros conglomerados, com aumento progressivo das distâncias.

f) Com base nos mesmos dados (sem as três maiores empresas), adotou-se agora a solução com seis conglomerados, novamente representada pelo gráfico Icicle. A

divisão em seis clusters permitiu uma segmentação mais refinada, revelando diferentes perfis de atuação empresarial. Havia conglomerados com redes locais pequenas, grupos em processo de expansão, empresas regionais consolidadas, entre outros.

g) Por fim, foi analisada a estrutura dos mesmos dados (sem outliers), mas considerando agora a divisão em três conglomerados, visualizada por meio do dendrograma. A solução revelou uma clara distinção entre empresas com estrutura reduzida (baixa área e poucos funcionários), empresas médias com boa distribuição regional e grupos maiores em crescimento. O dendrograma permitiu observar como as empresas foram se agrupando até chegar a essa estrutura, indicando afinidades estruturais claras entre os membros de cada cluster