DATA SCIENCE & ANALYTICS

MBA

CLUSTERING

0000010

Table of Contents

1. INTRODUÇÃO

Clusterização, ou agrupamento, é uma técnica de aprendizado não supervisionado utilizada em análise de dados para agrupar um conjunto de objetos de tal forma que os objetos no mesmo grupo (chamado de cluster) sejam mais similares entre si do que com aqueles em outros grupos (clusters).

1.1. OBJETIVO GERAL

O objetivo principal da análise de clusters é criar grupos que sejam:

- Homogêneos Internamente: Os dados dentro de cada cluster devem ser o mais semelhantes possível. Isso significa que a variabilidade dentro do cluster é minimizada, garantindo que os elementos do cluster compartilhem características comuns.

- Heterogêneos Entre Si: Os dados de diferentes clusters devem ser distintos uns dos outros. Isso significa que a variabilidade entre clusters é maximizada, destacando diferenças significativas entre os grupos.

1.2. TÉCNICA EXPLORATÓRIA (NÃO SUPERVISIONADA)

- A análise de agrupamentos caracteriza-se por ser uma técnica exploratória, de modo que não tem caráter preditivo para observações de fora da amostra;

- Se novas observações forem adicionadas à amostra, novos agrupamentos devem ser realizados, pois a inclusão de novas observações pode alterar a composição dos grupos;

- Se forem alteradas variáveis da análise, novos agrupamentos devem ser realizados, pois a inclusão/retirada de uma variável pode alterar os grupos;

2. TIPOS DE CLUSTERIZAÇÃO

2.1. CLUSTERIZAÇÃO PARTICIONAL:

- Divide o conjunto de dados em clusters distintos, onde cada ponto de dados pertence exatamente a um cluster. Exemplos incluem K-means e K-medoids.

    - K-means é um algoritmo de clusterização particional que particiona os dados em K clusters, onde K é um número pré-definido. O algoritmo minimiza a soma das distâncias quadráticas entre os pontos e os centróides dos clusters.

2.2. CLUSTERIZAÇÃO HIERÁRQUICA:

- Cria uma árvore de clusters, onde cada cluster pode ter sub-clusters. Esta abordagem pode ser
    - Aglomerativa: inicia com cada ponto como um cluster e vai unindo;
    - Divisiva: inicia com todos os pontos em um único cluster e vai dividindo;

➢ DENDOGRAMA: Métodos hierárquicos criam uma árvore de clusters com relações de hierarquia entre os clusters. O dendrograma é uma ferramenta comum para representar essa hierarquia.

2.3. CLUSTERIZAÇÃO BASEADA EM DENSIDADE:

- Forma clusters com base na densidade dos pontos de dados, como o DBSCAN, que identifica clusters como áreas de alta densidade de pontos separados por áreas de baixa densidade.

    - *DBSCAN (Density-Based Spatial Clustering of Applications with Noise):* algoritmo de clusterização baseado em densidade que identifica clusters como áreas densas de pontos, separadas por áreas de menor densidade. É útil para descobrir clusters de forma arbitrária e detectar outliers.

3. MÉTRICAS DE DISSIMILARIDADE

- Estas métricas são fundamentais para agrupar dados com base em suas similaridades.

- Uma métrica de dissimilaridade é uma função matemática usada para quantificar o grau de diferença ou similaridade entre dois ou mais objetos em um conjunto de dados.

No contexto da análise de dados e, especialmente, na clusterização, essas métricas são fundamentais para determinar a proximidade entre pontos de dados, influenciando diretamente como os clusters são formados.

- A maioria dos métodos de clusterização utiliza alguma forma de métrica de dissimilaridade para determinar a proximidade entre pontos de dados e formar clusters.

    - No entanto, existem alguns métodos de clusterização que não dependem explicitamente de métricas de dissimilaridade tradicionais. Um exemplo notável é o método de clusterização baseada em modos (mean-shift clustering).

- Diferentes métricas podem resultar em diferentes agrupamentos, afetando a estrutura e a interpretação dos clusters formados.

- Existem várias métricas de dissimilaridade, cada uma adequada para diferentes tipos de dados e objetivos analíticos. A maioria dos métodos de clusterização utiliza métricas de dissimilaridade para determinar a proximidade entre pontos de dados.

As Métricas de Dissimilaridade mais comuns incluem:

- Distância Euclidiana;
- Distância Manhattan;
- Distância de Minkowski;
- Distância de Mahalanobis;

4. ESQUEMAS DE AGLOMERAÇÃO

- Específico da Clusterização Hierárquica Aglomerativa;
- Esquemas de aglomeração referem-se aos processos de como os clusters são fundidos ao longo do tempo, resultando em uma hierarquia de clusters.
- Single Linkage, Complete Linkage, etc.

5. QUANTIDADE DE CLUSTERS

A escolha da quantidade de clusters é uma decisão crucial em qualquer análise de clusterização, pois determina como os dados serão agrupados e interpretados. Diferentes métodos de clusterização têm abordagens variadas para determinar o número ideal de clusters.

5.1. MÉTODOS PARA DETERMINAR QUANTIDADE DE CLUSTERS

5.1.1. MÉTODOS VISUAIS E HEURÍSTICOS:

- Método do Cotovelo (Elbow Method): Envolve plotar a soma dos erros quadrados (SSE) contra o número de clusters e procurar um ponto onde a redução no SSE começa a diminuir drasticamente. Esse ponto é chamado de "cotovelo" e sugere um bom número de clusters.
- Critério da Silhueta (Silhouette Method): Mede a coesão e separação dos clusters. Um valor de silhueta próximo de 1 indica que os pontos estão bem agrupados. Plotando os valores médios da silhueta para diferentes números de clusters pode ajudar a identificar o número ótimo de clusters.

- Gap Statistic: Compara a variabilidade dentro dos clusters com a esperada em uma referência nula (distribuição aleatória dos dados). O número de clusters onde a diferença (gap) é maior sugere o número ideal de clusters.

5.1.2. Critérios Estatísticos:

- AIC (Akaike Information Criterion) e BIC (Bayesian Information Criterion): Utilizados em modelos baseados em probabilidade, como Gaussian Mixture Models (GMM), para escolher o número de clusters que equilibra a complexidade do modelo e a qualidade do ajuste.

5.1.3. Métodos Baseados em Validação Cruzada:

- Dividir os dados em conjuntos de treinamento e teste, ajustar os modelos de clusterização no conjunto de treinamento e avaliar a performance no conjunto de teste.

5.2. Quantidade nos Tipos de Clusterização

5.2.1. Clusterização Hierárquica:

- Determinação da quantidade de clusters da clusterização hierárquica é feita após a análise, usando um dendrograma.

    - Clusterização Hierárquica Aglomerativa (HAC):
        - Começa com cada observação em seu próprio cluster e, iterativamente, une os clusters mais próximos até que todos os pontos estejam em um único cluster.
        - Métodos de Encadeamento: Single Linkage, Complete Linkage, Average Linkage, Ward's Method.

- Clusterização Hierárquica Divisiva:
    - Começa com todos os pontos em um único cluster e divide-os iterativamente até que cada observação esteja em seu próprio cluster.

5.2.2. Clusterização Particional

- Determinação da quantidade de clusters da clusterização particional é feita usando métodos como o Elbow Method, Critério da Silhueta ou Gap Statistic.

    - K-means:
        - Particiona os dados em K clusters, onde K é especificado a priori. O algoritmo ajusta iterativamente os centróides dos clusters e reatribui pontos até que a variância dentro dos clusters seja minimizada.

    - K-medoids:
        - Semelhante ao K-means, mas usa pontos reais dos dados como centróides (medoids), tornando-o mais robusto a outliers.

5.2.3. Clusterização Baseada em Densidade

- A determinação da quantidade de clusters da clusterização baseada em densidade é automática.

    - DBSCAN (Density-Based Spatial Clustering of Applications with Noise):
        - Forma clusters com base na densidade de pontos. Identifica clusters como regiões densas separadas por regiões de menor densidade.

    - OPTICS (Ordering Points To Identify the Clustering Structure):
        - Similar ao DBSCAN, mas pode identificar clusters de densidade variável.

5.2.4. Clusterização Baseada em Modelos

- A determinação da quantidade de clusters para clusterização baseada em modelos é feita usando critérios como AIC ou BIC.

    - Gaussian Mixture Models (GMM):
        - Assume que os dados são gerados a partir de uma mistura de várias distribuições Gaussianas. Estima os parâmetros dessas distribuições para formar clusters.

5.2.5. Clusterização Baseada em Grafos

- A determinação da quantidade de clusters é feita usando métodos como o Elbow Method ou a análise do espectro da matriz de afinidade.

    - Spectral Clustering:
        - Utiliza a teoria dos grafos e a decomposição espectral da matriz de afinidade para realizar a clusterização. Transforma o problema de clusterização em um problema de partição de grafos.

6. MÉTODOS

Analisaremos dois métodos para a obtenção de agrupamentos

- Método Hierárquico Aglomerativo:
    - A quantidade de clusters é definida ao longo da análise (passo a passo);
- Método Não Hierárquico K-means
    - Define-se a priori quantos cluster serão formados;

6.1. MÉTODO HIERÁRQUICO AGLOMERATIVO

Na análise hierárquica aglomerativa, a análise começa com cada observação sendo tratada como um cluster separado, e esses clusters são então combinados sucessivamente com base em uma medida de dissimilaridade e um método de encadeamento escolhidos.

6.1.1. TRATAMENTO INICIAL DOS DADOS

Análise das variáveis que serão estudadas:

- Antes de iniciar os procedimentos, é importante realizar uma análise das unidades de medidas das variáveis;
- Se estiverem em unidades de medidas distintas, é importante realizar a padronização das variáveis antes de iniciar a análise de cluster;
- Aplica-se o ZScore (torna variáveis com média = 0 e desvio padrão = 1);

6.1.2. Escolhas inerentes ao método

Na análise de cluster hierárquica aglomerativa, é necessário escolher uma medida de dissimilaridade e um método de encadeamento. Essas escolhas afetam diretamente como os clusters são formados e unidos.

- Escolha da medida de dissimilaridade (distância);
  - por exemplo, distância Euclidiana, Manhattan, etc.
  - Refere-se à distância entre as observações, com base nas variáveis escolhidas;
  - Portanto, indica o quanto as observações são diferentes entre si;

- Escolha do método de encadeamento das observações
  - por exemplo, single linkage, complete linkage, average linkage, etc.
  - Refere-se à especificação da medida de distância quando houver cluster formados

6.1.3. Esquemas de aglomeração

Clusterização Hierárquica Aglomerativa: Este termo refere-se ao processo de como os clusters são fundidos ao longo do tempo, resultando em uma hierarquia de clusters. É um componente central deste método.

Hierárquico aglomerativo: observações separadas → um único cluster

- Considerando n observações, inicia-se com n clusters (estágio 0);

- Na sequência, une-se as duas observações com menor distância (n-1 clusters);

- Em seguida, um novo grupo é formado pela união de duas novas observações ou pela inclusão de uma observação ao cluster formado na etapa anterior (sempre pela menor distância). O método de encadeamento indica qual é a distância a ser considerada;

- Repete-se a etapa anterior n-1 vezes, ou seja, até restar somente 1 cluster;

- O dendrograma é um gráfico que permite visualizar a formação dos clusters;

6.1.4. Medidas de dissimilaridade

As medidas de dissimilaridade são fundamentais para a análise de clusterização, especialmente em métodos hierárquicos. Elas quantificam a diferença ou similaridade entre pontos de dados, servindo como base para a formação de clusters. Diferentes

medidas de dissimilaridade podem levar a diferentes agrupamentos, por isso a escolha adequada da medida é crucial para obter resultados significativos e interpretáveis.

A escolha da medida de dissimilaridade (como distância Euclidiana, Manhattan, Minkowski, etc.) pode impactar significativamente o resultado da análise de clustering hierárquico, levando a diferentes estruturas de clusters.

Na clusterização hierárquica, as medidas de dissimilaridade desempenham um papel crucial na determinação de como os clusters são formados e unidos. O processo de clusterização envolve a fusão iterativa de clusters baseados na menor medida de dissimilaridade, conforme definido pelo método de encadeamento escolhido (single linkage, complete linkage, average linkage, etc.). A escolha da medida de dissimilaridade pode influenciar significativamente a estrutura e a interpretação dos clusters resultantes.

Principais Medidas de Dissimilaridade

- Distância Euclidiana:
    - Descrição: Mede a distância direta entre dois pontos no espaço n-dimensional. É a medida mais comum e intuitiva, baseada no teorema de Pitágoras.
    - Aplicação: Utilizada quando as variáveis são contínuas e escalas similares. Ideal para dados em que a proximidade física tem significado.

- Distância de Manhattan:
    - Descrição: Calcula a soma das diferenças absolutas entre as coordenadas dos pontos. Representa a distância percorrida em um grid, como ruas de uma cidade.
    - Aplicação: Preferida em contextos onde os deslocamentos são restritos a eixos principais (e.g., distâncias em ruas de uma cidade).
- Distância de Minkowski:
    - Descrição: Generalização das distâncias Euclidiana (p=2) e Manhattan (p=1). Permite ajustar a medida de dissimilaridade com o parâmetro ppp.
    - Aplicação: Versátil, pode ser ajustada para diferentes necessidades e propriedades dos dados.
- Distância de Chebyshev:
    - Descrição: Considera a maior diferença absoluta entre as coordenadas dos pontos. É a distância máxima em qualquer dimensão.
    - Aplicação: Útil quando o critério de dissimilaridade é dominado pela maior diferença em uma dimensão.
- Distância de Mahalanobis:
    - Descrição: Leva em conta a correlação entre variáveis, ajustando a escala das diferenças. SSS é a matriz de covariância dos dados.
    - Aplicação: Ideal para dados onde as variáveis são correlacionadas, proporcionando uma medida de dissimilaridade normalizada.

- Coeficiente de Correlação
    - Descrição: Mede a relação linear entre duas variáveis. Varia de -1 a 1, onde 1 indica correlação perfeita, -1 indica correlação inversa perfeita, e 0 indica ausência de correlação.
    - Aplicação: Utilizada em análise de dados multivariados, onde a relação entre variáveis é mais informativa que a distância absoluta.

- Identifica a distância entre observações
  - Distância de Minkowski: $d_{pq} = \left[\sum_{j=1}^{k}(|ZX_{jp} - ZX_{jq}|)^m\right]^{\frac{1}{m}}$
  - Distância euclidiana: $d_{pq} = \sqrt{\sum_{j=1}^{k}(ZX_{jp} - ZX_{jq})^2}$
  - Distância euclidiana quadrática: $d_{pq} = \sum_{j=1}^{k}(ZX_{jp} - ZX_{jq})^2$
- Distância de Manhattan (City Block): $d_{pq} = \sum_{j=1}^{k}|ZX_{jp} - ZX_{jq}|$
- Distância de Chebychev: $d_{pq} = max\,|ZX_{jp} - ZX_{jq}|$
- Distância de Canberra: $d_{pq} = \sum_{j=1}^{k}\frac{|ZX_{jp} - ZX_{jq}|}{(ZX_{jp} + ZX_{jq})}$ → variáveis de valores positivos
- A correlação de Pearson entre as observações também pode ser utilizada (mas é uma medida de semelhança, portanto ajusta-se sua interpretação)

6.1.5. MÉTODOS DE ENCADEAMENTO

Métodos de Encadeamento definem como as distâncias entre clusters são calculadas e como os clusters são fundidos ao longo do processo de aglomeração.

- Indica qual distância utilizar quando já existem clusters formados durante os estágios aglomerativos;

- específicos da clusterização hierárquica, especialmente da aglomerativa.

O método de encadeamento influencia como as distâncias entre clusters são calculadas e, portanto, pode levar a diferentes agrupamentos hierárquicos.

Método de Encadeamento	Ilustração	Distância (Dissimilaridade)
Único *(Nearest Neighbor ou Single Linkage)*		d_{23}
Completo *(Furthest Neighbor ou Complete Linkage)*		d_{15}
Médio *(Between Groups ou Average Linkage)*		$\frac{d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}}{6}$

6.1.5.1. NEAREST NEIGHBOR (SINGLE LINKAGE):

- privilegia menores distâncias, recomendável em casos de observações distintas;

- Tende a formar clusters alongados ou em cadeia, pois é sensível a outliers.

- Em casos de observações distintas, onde os clusters podem ter formas alongadas ou irregulares, o Single Linkage é útil porque pode capturar a estrutura dos dados sem forçar os clusters a se ajustarem a formas predefinidas.

- Single Linkage consegue unir pontos de dados em clusters que seguem a distribuição natural dos dados, mesmo que sejam alongados ou curvados.

- Identificação de Formas Não Convexas: O Single Linkage é eficaz na identificação de formas não convexas porque ele não impõe a restrição de que os clusters devem ser esféricos ou convexos. Clusters podem se estender de maneiras que seguem a distribuição natural dos dados.

- Formas não convexas: são formas onde existem pares de pontos que, ao traçar uma linha reta entre eles, a linha sai da forma. Exemplos incluem formas em "U", "L" ou anéis.

- Bom para Observações Distintas: situações onde os dados possuem uma grande variabilidade e onde os clusters podem ter formas irregulares e não uniformes;

6.1.5.2. Between groups (average linkage):

- Junção de grupos pela distância média entre todos os pares de observações do grupo em análise;

- Balanceia entre a formação de clusters alongados e compactos, sendo um compromisso entre single e complete linkage.

6.1.5.3. Furthest neighbor (complete linkage):

Privilegia maiores distâncias, recomendável em casos de observações parecidas;
O método "complete linkage" considera a distância máxima entre quaisquer pares de observações, onde cada par é composto de uma observação de cada cluster. Em outras palavras, é a maior distância entre pontos que pertencem a clusters diferentes.

- Formação de Clusters Compactos: Devido ao critério de maior distância, o Complete Linkage tende a formar clusters mais compactos e globulares;

- Clusters Homogêneos: A maior distância entre quaisquer dois pontos é usada para unir clusters, o que evita a formação de clusters alongados ou "cadeias".

- Menos Sensível a Outliers: Ao considerar a maior distância, Complete Linkage é menos sensível a outliers em comparação com Single Linkage. Isso significa que um ponto distante não afetará significativamente a união dos clusters.

- Aplicabilidade a dados de alta variabilidade;

- Variabilidade Reduzida Dentro dos Clusters: Complete Linkage é adequado para dados onde se deseja minimizar a variabilidade dentro dos clusters, garantindo que os pontos dentro de um cluster sejam relativamente próximos uns dos outros.

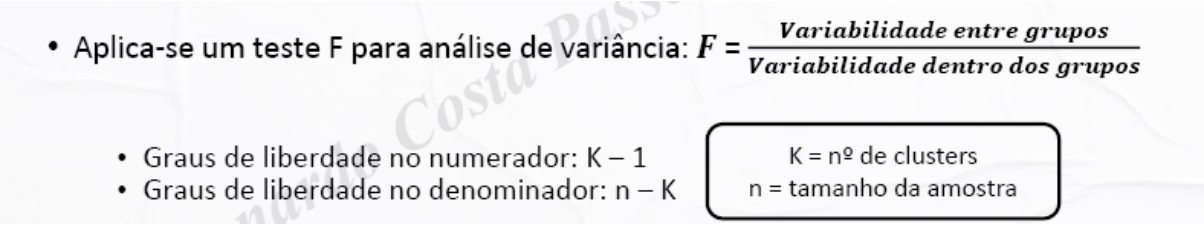
6.1.6. Quantos clusters escolher?

- Como critério para a escolha do número final de clusters em uma análise, pode-se adotar o tamanho dos saltos de distância para a incorporação seguinte;
- Saltos muito elevados podem indicar o agrupamento de observações com características mais distintas, isto é, há a união de observações mais distintas;
- Comparar dendrogramas obtidos por diferentes métodos de encadeamento: dendrograma é um gráfico de árvore que mostra a disposição dos clusters em cada nível de aglomeração, mostrando a hierarquia de fusão dos clusters.

6.1.7. Análise dos agrupamentos

Quais variáveis contribuem?

- Após a clusterização, é importante comparar se a variabilidade dentro do grupo é menor do que a variabilidade entre grupos com base nas variáveis da análise



- É possível analisar quais variáveis mais contribuíram para a formação de pelo menos um dos clusters: maiores valores da estatística F (em conjunto com a significância);

6.2. Método Não Hierárquico K-means

6.2.1. Tratamento inicial dos dados

Análise das variáveis que serão estudadas:

- Também é importante realizar a análise das unidades de medidas das variáveis para a aplicação do K-means;
- Se estiverem em unidades de medidas distintas, é fundamental padronizar as variáveis antes de iniciar a análise;
- Padronização pelo ZScore (variáveis com média = 0 e desvio padrão = 1): a padronização por meio do ZScore é necessária quando as variáveis têm diferentes unidades de medida para garantir que cada variável contribua igualmente para a métrica de dissimilaridade.

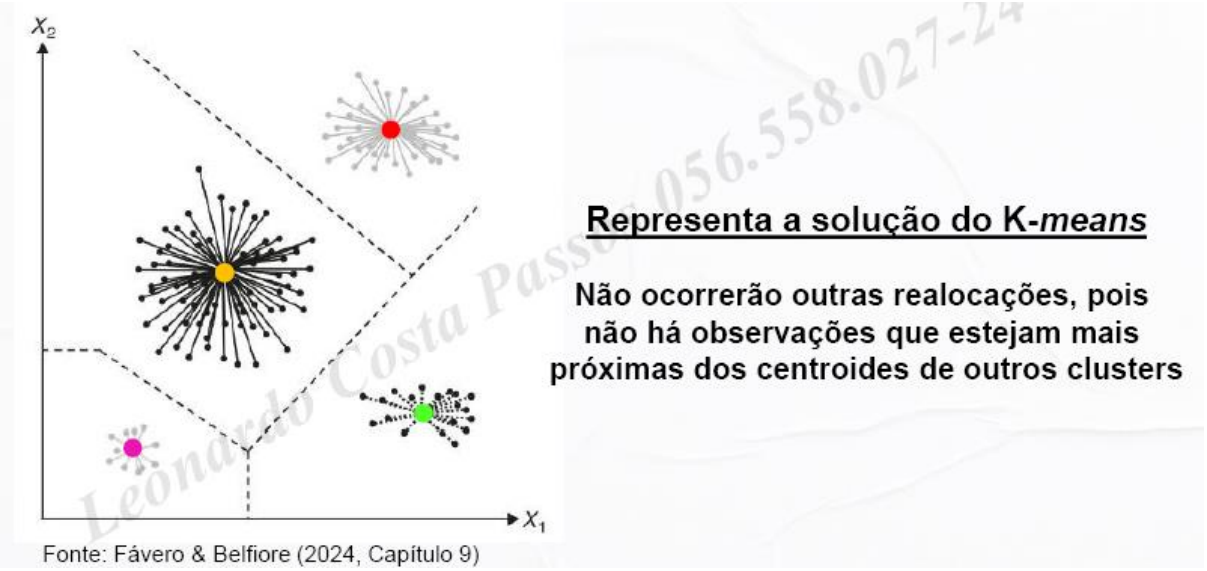
6.2.2. Esquemas de aglomeração

- A quantidade K de clusters é escolhida a priori e é usada como base para a identificação dos centros de aglomeração, de modo que as observações são arbitrariamente alocadas aos K clusters para o cálculo dos centroides iniciais;

- Nas etapas seguintes, as observações vão sendo comparadas pela proximidade aos centroides dos outros clusters. Se houver realocação a outro cluster por estar mais próxima, os centroides são recalculados (em ambos os clusters);

- Trata-se de um processo iterativo;

- O procedimento K-means encerra-se quando não for possível realocar qualquer observação por estar mais próxima do centroide de outro cluster: indica que a soma dos quadrados de cada observação até o centro do cluster alocada foi minimizada;



Fonte: Fávero & Belfiore (2024, Capítulo 9)

6.2.3. Identificação da quantidade de clusters

Técnicas para a identificação da quantidade de clusters no K-means

- Método de Elbow: calcula-se a soma total dos quadrados dentro dos clusters (WCSS) para várias opções de K (quantidade de clusters). No gráfico, busca-se a dobra (“cotovelo”), ou seja, o ponto a partir do qual a

diminuição na WCSS não é mais tão expressiva, mesmo aumentando a quantidade de clusters;

- **Método da Silhueta**: para cada observação, calcula-se: **(b)** sua distância média para o cluster mais próximo onde não esteja alocada; **(a)** sua distância média dentro do cluster onde está alocada

$$\textbf{silhueta} = \frac{(\mathbf{b}-\mathbf{a})}{\mathbf{max}(\mathbf{a},\mathbf{b})}$$

Quanto mais próximo de 1, melhor a clusterização. Quanto mais próximo de -1, pior!

- Em seguida, calcula-se o **coeficiente de silhueta médio** para todas as observações. O procedimento é realizado para várias opções de K

7. CONSIDERAÇÕES

- Quando há variáveis categóricas, pode aplicar a Análise de Correspondência;
- O output do método hierárquico pode ser utilizado como input no método não hierárquico para a identificação inicial da quantidade de clusters;
- O método não hierárquico k-means pode ser aplicado em amostras maiores;