UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

Calvin Rodrigues

Técnicas e visualização de agrupamentos em aprendizagem não supervisionada com aplicações

Juiz de fora
2021

Calvin Rodrigues

Técnicas e visualização de agrupamentos em aprendizagem não supervisionada com aplicações

Trabalho de Conclusão de Curso apresentado ao Curso de Estatística da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do título de Bacharel em Estatística.

Orientador: Dr. Lupércio França Bessegato

Juiz de Fora
2021

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Rodrigues, Calvin.
Técnicas e visualização de agrupamentos em aprendizagem não supervisionada com aplicações / Calvin Rodrigues. -- 2021.
58 f. : il.

Orientador: Lupércio França Bessegato
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas, 2021.

1. Aprendizagem não supervisionada. 2. Análise de agrupamento. 3. Redução de dimensionalidade. 4. Escalonamento multidimensional. 5. Visualização multivariada. I. Bessegato, Lupércio França, orient. II. Título.

Calvin Rodrigues

Técnicas e visualização de agrupamentos em aprendizagem não supervisionada com aplicações

Trabalho de Conclusão de Curso apresentado ao Curso de Estatística da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do título de Bacharel em Estatística.

Aprovado em 10 de setembro de 2021

BANCA EXAMINADORA

_____

Professor Dr. Lupércio França Bessegato – Orientador

Universidade Federal de Juiz de Fora

_____

Professora Dra. Camila Borelli Zeller

Universidade Federal de Juiz de Fora

_____

Professor Dr. Gustavo de Carvalho Lana

Universidade Federal de Juiz de Fora

AGRADECIMENTOS

Gostaria de agradecer aos meus pais, Heloisa e Ivanir, meus irmãos, Kaian, Caio e Maria Eduarda, meu namorado Victor Hugo e meus amigos, em especial os que conheci durante a graduação, que sempre me apoiaram, estimularam e incentivaram a acreditar no meu potencial.

Agradeço também aos professores do departamento de estatística que contribuíram para meu aprendizado e formação acadêmica, em especial ao meu orientador, Lupércio Bessegato por todo suporte e incentivo.

RESUMO

Na última década, a quantidade de informações armazenadas no formato digital tem crescido exponencialmente, levando à necessidade cada vez maior de produção de procedimentos computacionalmente intensivos que auxiliem na geração de informação a partir desses dados. Dentre outras, a área de aprendizagem estatística não supervisionada fornece técnicas que auxiliam na obtenção de informação a partir desses dados sem que estejam associados a respostas. Dessa maneira, esse trabalho propôs estudar técnicas de agrupamento e de redução de dimensionalidade, a fim de explicar a configuração dos dados a partir de medida de distância entre objetos. Em redução de dimensionalidade foram explorados métodos de escalonamento multidimensional métrico e não métrico para visualizar previamente os possíveis grupos formados em um espaço de dimensão reduzida. Por sua vez, para agrupamento, foram utilizados os procedimentos K-médias, AGNES e DIANA, no qual o primeiro, que agrupa de maneira particionada, solicita previamente o número de grupos a ser formado, enquanto os demais, que agrupam hierarquicamente, contornam esse problema em troca da definição da medida de ligação entre grupos. Por fim, as metodologias estudadas foram aplicadas em conjuntos de dados reais com o *software* R Core Team (2020).

Palavras-chave: Aprendizagem não supervisionada. Análise de agrupamento. Redução de dimensionalidade. Escalonamento multidimensional. Visualização multivariada.

ABSTRACT

In the last decade, the amount of information stored in digital format has grown exponentially, leading to the growing need to produce computationally intensive procedures that help generate information from this data. Among others, the area of unsupervised statistical learning provides techniques that help to obtain information from these data without being associated with answers. Thus, this work proposed to study clustering and dimensionality reduction techniques, in order to explain the data configuration based on measuring the distance between objects. In dimensionality reduction, metric and non-metric multidimensional scaling methods were explored to preview the possible groups formed in a reduced dimension space. In turn, for clustering, the K-means, AGNES and DIANA procedures were used. The first, which groups in a partitioned manner, previously requests the number of groups to be formed, while the others, which group hierarchically, bypass this problem in exchange for defining the measure of linkage between groups. Finally, the studied methodologies were applied to real datasets with the R Core Team software (2020).

Keywords: Unsupervised learning. Cluster analysis. Dimensionality reduction. Multidimensional scaling. Multivariate visualization.

LISTA DE ILUSTRAÇÕES

Figura 1 cMDS com uma dimensão 30
Figura 2 cMDS com duas dimensões 30
Figura 3 Métodos silhueta média e estatística de lacuna para escolha do número de grupos 31
Figura 4 Métodos cotovelo (wss) e CCC para escolha do número de grupos 32
Figura 5 Agrupamento K-médias utilizando cMDS com uma dimensão 33
Figura 6 Agrupamento K-médias utilizando cMDS com duas dimensões 33
Figura 7 Método hierárquico aglomerativo com variação mínima de Ward 35
Figura 8 Método hierárquico divisivo 35
Figura 9 nMDS com duas dimensões 36
Figura 10 Métodos cotovelo (wss), silhueta média e estatística de lacuna para determinação do número de grupos 37
Figura 11 Critério CCC para determinação do número de grupos 38
Figura 12 AGNES com variação mínima de Ward 39
Figura 13 DIANA 40

LISTA DE TABELAS

Tabela 1 Critério Stress 19
Tabela 2 Coeficiente R^2 para diferentes medidas de dissimilaridade 31
Tabela 3 Resumo das variáveis originais por cada grupo dado por K-médias 32
Tabela 4 Coeficiente aglomerativo para quatro métodos de ligação 34
Tabela 5 Coeficiente aglomerativo para diferentes métodos de ligação 38
Tabela 6 Resumo das variáveis originais em cada grupo dado por AGNES ... 39
Tabela 7 Resumo das variáveis originais em cada grupo dado por DIANA 40

SUMÁRIO

1 INTRODUÇÃO 12
2 MEDIDAS DE DISTÂNCIA ENTRE OBSERVAÇÕES 13
2.1 DISTÂNCIA EUCLIDIANA 14
2.2 DISTÂNCIA DE MANHATTAN 14
2.3 DISTÂNCIA DE MINKOWSKI 15
2.4 DISTÂNCIA DE MAHALANOBIS 15
2.5 DISTÂNCIA DE GOWER 16
3 ESCALONAMENTO MULTIDIMENSIONAL 17
3.1 ESCALONAMENTO MULTIDIMENSIONAL CLÁSSICO (cMDS) 18
3.2 ESCALONAMENTO MULTIDIMENSIONAL NÃO MÉTRICO (nMDS) 18
3.3 QUALIDADE DO AJUSTE 19
3.3.1 Coeficiente de mensuração 19
3.3.2 Stress 19
4 ANÁLISE DE AGRUPAMENTO 20
4.1 MÉTODOS DE PARTICIONAMENTO 20
4.1.1 Agrupamento por K-médias 20
4.1.1.1 Algoritmo de Hartigan-Wong 21
4.1.2 Qualidade do agrupamento 22
4.2 MÉTODOS HIERÁRQUICOS 22
4.2.1 Métodos de ligação 23
4.2.1.1 Ligação única ou mínima 23
4.2.1.2 Ligação completa ou máxima 24
4.2.1.3 Ligação média 24
4.2.1.4 Método de variação mínima de Ward 24
4.2.2 Avaliação da estrutura hierárquica 25
5 DETERMINAÇÃO DO NUMERO IDEAL DE GRUPOS 26
5.1 MÉTODO COTOVELO 26
5.2 MÉTODO DE SILHUETA MÉDIA 27
5.3 CUBIC CLUSTERING CRITERION (CCC) 27
5.4 MÉTODO DE ESTATÍSTICA DE LACUNA 28
6 APLICAÇÕES 29
6.1 IRIS 29
6.1.1 Redução de dimensionalidade 29
6.1.2 Análise de agrupamento 29
6.2 SEGMENTAÇÃO DE CLIENTES 36

6.2.1 Redução de dimensionalidade 36
6.2.2 Análise de agrupamento 37
7 CONSIDERAÇÕES FINAIS 41
REFERÊNCIAS 42
APÊNDICE 44

1 INTRODUÇÃO

A tentativa de simular a característica humana de absorver e assimilar conhecimento em computadores caracteriza a área de pesquisa chamada aprendizado de máquina (DUDA; STORK; HART, 2000). Ela procura desenvolver métodos computacionais capazes de aprender com experiências passadas (BISHOP, 2006).

As técnicas de aprendizado de máquina podem gerar modelos capazes de reconhecer e imitar o comportamento humano de modo automático. Essas técnicas são divididas, de maneira geral, em duas abordagens: aprendizagem supervisionada e não supervisionada.

Em aprendizagem supervisionada, ou aprendizagem a partir de exemplos, a máquina, além de receber os dados de entrada $x_1, x_2, \dots, x_n$, recebe uma sequência de respostas (saídas) desejadas $y_1, y_2, \dots, y_n$ e seu papel é aprender através desses conjuntos para conseguir retornar a verdadeira resposta a cada nova entrada (MARSLAND, 2014).

A classificação e a regressão são os problemas mais comumente encontrados em aprendizagem supervisionada. No primeiro, os valores das classes são discretos, já no segundo problema, as classes assumem valores contínuo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Por outro lado, em aprendizagem não supervisionada, temos os dados de entrada, porém, não estamos interessados na predição, já que as variáveis respostas associadas às entradas não serão dadas. Em vez disso, o objetivo é explicar a configuração dos dados através de representações (JAMES *et al.*, 2013).

Muitos autores consideram agrupamento de dados e redução de dimensionalidade, ou *clustering,* as principais técnicas em aprendizagem estatística não supervisionada. Os métodos de agrupamento consistem em encontrar grupos em determinado conjunto de dados nos quais os itens dentro de um grupo sejam os mais similares possíveis entre si, enquanto que os itens em grupos diferentes sejam os mais distintos possíveis (BISHOP, 2006; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; MARSLAND, 2014). Já os métodos de redução de dimensionalidade introduzem um novo espaço, obrigatoriamente de menor dimensão que o espaço original, no qual os objetos (valores das variáveis) são representados.

No que se diz respeito às técnicas mencionadas acima, Hastie et al. (2009) reforçam que não há uma mensuração direta de sucesso, o que leva a uma forte proliferação de métodos propostos, uma vez que a eficácia é uma questão que, no geral, não pode ser verificada diretamente. Todo conjunto de dados pode ser agrupado, porém, para que o agrupamento seja o melhor possível é necessário que se conheça principalmente as particularidades das medidas de dissimilaridade.

Neste trabalho serão estudados métodos de agrupamento particionado e hierárquico, além de técnicas de escalonamento multidimensional para redução de dimensionalidade a partir da visualização da matriz de distâncias entre observações, a fim de explicar a configuração dos dados a partir de determinada medida de distância entre objetos.

2 MEDIDAS DE DISTÂNCIA ENTRE OBSERVAÇÕES

Hastie et al. (2009) consideram fundamental para todas as técnicas de agrupamento a escolha da medida de distância, ou dissimilaridade, entre dois objetos. Naturalmente, quando trabalhando com agrupamento, estamos assumindo que todos os relacionamentos podem ser descritos por uma matriz contendo essas medidas de dissimilaridade, ou de proximidade, entre cada par de observações para que, com ela, possamos criar grupos mais homogêneos possível.

Cada entrada d_{ij} na matriz consiste em um valor numérico que demonstra quão distantes os valores *i* e *j* são. É importante ressaltar que algumas métricas calculam a dissimilaridade (distância) e outras a similaridade (correlação), os mais populares algoritmos de agrupamento utilizam a matriz de dissimilaridade, porém, em ambos os casos a essência é a mesma (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Todos os coeficientes de dissimilaridade são funções $d: \Gamma \, X \, \Gamma \Rightarrow \mathbb{R}$, em que Γ denota o conjunto de objetos com o qual estamos trabalhando, essas funções permitem realizar a transformação da matriz de dados (2.1) em uma matriz de distâncias (2.2).

$$\Gamma = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}_{nXp} \tag{2.1}$$

$$\mathrm{D} = \begin{bmatrix} 0 & & & \\ d_{2,1} & 0 & & \\ \vdots & \vdots & \ddots & \\ d_{n,1} & d_{n,2} & \cdots & 0 \end{bmatrix}_{nXn} \tag{2.2}$$

Todas as funções de dissimilaridade têm como critérios básicos os seguintes:

- $d_{ii} = 0, \forall\, i \in \Gamma$
- $d_{ij} \geq 0, \ \forall\, i, j \in \Gamma$
- $d_{ij} = d_{ji}, \ \forall\, i, j \in \Gamma$. Esta regra afirma que a distância entre dois elementos não varia independente do ponto a partir do qual ela é medida. Por isso 2.2 é mostrada sendo triangular inferior. Caso a matriz original D não for simétrica, ela deve ser substituída por $(D + D^T)/2$ (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).
- $d_{ij} + d_{jg} \geq d_{ig}, \ \forall\, i, j, g \in \Gamma$. Se trata da desigualdade triangular e especifica que a menor distância entre dois pontos é uma reta.

Se, além das propriedades citadas acima, a métrica também possuir a propriedade $d(ax_i, ax_j) = |a|\, d(x_i, x_j)$, então ela é chamada de uma norma.

As medidas de dissimilaridade são utilizadas para calcular a distância entre cada par de objetos *i* e *j*, objeto *i* caracterizado pelos atributos $x_i = (x_{i1}, \dots, x_{ip})$ e objeto *j* caracterizado pelos atributos $x_j = (x_{j1}, \dots, x_{jp})$, na matriz de dados (2.1), sendo *p* o número de variáveis estudadas (ou número de colunas).

Dentre as métricas para cálculo de distância entre objetos iremos nos ater as medidas: Euclidiana, Manhattan, Minkowski, Mahalanobis e Gower.

2.1 DISTÂNCIA EUCLIDIANA

Para os mais comuns softwares de agrupamento, a métrica de distância padrão é a euclidiana (KASSAMBARA, 2017). Essa distância é dada pela fórmula pitagórica abaixo que resulta na reta que liga os dois pontos em questão. Quanto menor o valor da distância entre dois objetos, mais parecidos eles vão ser e, portanto, serão agrupados em um mesmo cluster.

Apesar de muito comum, a distância euclidiana não pode ser utilizada para dados não métricos.

$$d_E(i,j) = \sqrt{\sum_{g=1}^{p} \left(x_{ig} - x_{jg}\right)^2}, \qquad i \neq j \tag{2.3}$$

Em termos relativos a utilização a distância euclidiana se aplica melhor a dados não padronizados, ou seja, dados que não têm nenhum tipo de tratamento de adaptação de escala, fazendo com que o resultado final seja insensível a valores discrepantes.

Uma desvantagem sobre essa medida de distância pode acontecer se houver diferença de escala entre as dimensões; por exemplo, se um objeto (ou variável) for medido em quilômetros e outro em centímetros, no momento em que houver a transformação de escala as distâncias medidas sofrerão uma fortíssima influência daquele objeto com os maiores valores.

2.2 DISTÂNCIA DE MANHATTAN

O cálculo da distância de Manhattan é dado através da soma das diferenças absolutas dos atributos x_i e x_j, respectivos aos objetos *i* e *j*, como dada abaixo

$$d_M(i,j) = \sum_{g=1}^{p} \left|\left(x_{ig} - x_{jg}\right)\right|, \qquad i \neq j \tag{2.4}$$

Assim como a distância euclidiana, a métrica de Manhattan é insensível a valores discrepantes, em contrapartida não há influência de escala sobre o resultado, visto que os valores não são elevados ao quadrado como na distância anterior.

Com isso, em conjuntos de dados métricos cujos objetos têm alta diferença de escala, a distância de Manhattan se destaca em eficácia quando comparada com a distância euclidiana.

2.3 DISTÂNCIA DE MINKOWSKI

A distância de Minkowski é uma métrica considerada como uma generalização de ambas as distâncias euclidiana e Manhattan e é dada pela seguinte equação:

$$d_{Mi}(i,j) = \sum_{g=1}^{p} \sqrt[q]{|x_{ig} - x_{jg}|^q} \tag{2.5}$$

Podemos ver que as métricas de Minkowski têm a propriedade de norma. A escolha do valor de q depende única e exclusivamente da ênfase que se deseja dar a distâncias maiores. Quanto maior seu valor, maior a sensitividade da métrica a essas distâncias.

2.4 DISTÂNCIA DE MAHALANOBIS

A distância de Mahalanobis difere das anteriormente vistas por levar em consideração a correlação entre os vetores. O cálculo da distância entre os objetos i e j da mesma distribuição que possuam uma matriz de covariância Σ é dado por

$$d_{Ma}(i,j) = \sqrt{(\overrightarrow{x_i} - \overrightarrow{x_j})^T \, \Sigma^{-1} \, (\overrightarrow{x_i} - \overrightarrow{x_j})} \tag{2.6}$$

Nota-se que se a matriz de covariância é a matriz identidade, 2.6 se reduz à distância euclidiana.

A maneira comum de lidar com a correlação e ponderação diferencial de variáveis em muitas técnicas multivariadas é através da distância de Mahalanobis, porém, apesar de ser muito utilizada no ambiente multivariado, essa medida raramente é adotada para fins de agrupamento, a principal razão para qual é que a definição dessa distância é para objetos de uma única população.

Além disso, as matrizes de covariância podem ser difíceis de determinar e a memória e o tempo de computação crescem de forma quadrática com o número de características.

Uma possível vantagem do uso dessa métrica é a detecção de valores discrepantes, pois um valor alto determina que certa observação está a vários desvios padrões do centro. Se compararmos com outros grupos existentes, cujas distribuições sejam conhecidas, essa métrica pode ser usada como maneira de determinar a qual grupo uma observação deve pertencer.

2.5 DISTÂNCIA DE GOWER

Uma popular métrica de distância híbrida (para lidar com dados métricos e não métricos simultaneamente) é a distância de Gower (GOWER, 1971). A distância entre os dois objetos *i* e *j* de tamanho *p* é dada a seguir.

Considere $f_g\,(i;j) \coloneqq \left|x_{ig} - x_{jg}\right|/r_g$ quando temos objetos métricos intervalares, com r_g representando a amplitude da variável *g*, e $f_g\,(i;j) \coloneqq I\left\{x_{ig} \neq x_{jg}\right\}$ se os objetos são não métricos, sendo I a função indicadora. Então, a distância de Gower é dada por:

$$d_G(i,j) = \frac{\sum_{g=1}^{p} w_g \,.\, f_g(i,j)}{\sum_{g=1}^{p} w_g}, \tag{2.7}$$

na qual w_g é um peso especificado pelo usuário para as variáveis *g* = 1, 2, ..., *p*, a maioria dos softwares utiliza peso igual a 1 como padrão.

Foss et al. (2019) consideram a escolha efetiva dos pesos de cada variável uma tarefa difícil devido a falta de referência, ou seja, uma configuração que funciona para um cenário não é garantia que funcione para outros, mesmo que sejam próximos.

Além disso, a função de distância categórica utilizada no cálculo da distância de Gower é insensível às probabilidades de cada nível categórico dentro dos clusters, o que pode limitar sua eficiência quando utilizada em dados não métricos (FOSS; MARKATOU; RAY, 2019).

A vantagem do uso dessa distância sobre as outras estudadas até então se dá no uso de dados híbridos, visto que as distâncias vistas anteriormente só podem ser utilizadas em dados métricos.

3 ESCALONAMENTO MULTIDIMENSIONAL

Dado um conjunto de objetos tais que seus atributos individuais são coordenadas de pontos de um espaço com mais de duas dimensões, as técnicas de escalonamento multidimensional permitem representar esses objetos como pontos de um novo espaço de dimensão reduzida, tornando visualizável as dissimilaridades entre os objetos do conjunto de dados em questão (BORG; GROENEN, 2005).

O escalonamento multidimensional (MDS) é considerado como um modelo espacial para representação da matriz de dissimilaridades (2.2), apesar de ser essencialmente uma técnica de redução de dimensionalidade e pode ser muito eficaz visualmente para apontar grupos em um conjunto de dados. A grande vantagem do uso do MDS consiste no fato de que é suficiente conhecermos apenas as distâncias, ou dissimilaridades, entre os objetos no espaço original, sem que as coordenadas de cada ponto sejam especificadas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Conhecendo-se as distâncias d_{ij} entre os objetos do conjunto de dados, o MDS determina pontos $z_1, \dots, z_n \in \mathbb{R}^q$, em que q, escolhido pelo usuário, é menor que o número de dimensões do conjunto original. Esses pontos são criados tal que a distância entre um par de pontos é aproximadamente igual à distância entre dois objetos originais.

Em geral, mas não exclusivamente, assume-se que as distâncias entre os pontos no modelo espacial são euclidianas, $d_{ij} = \|x_i - x_j\|$. Hastie et al. (2009) definem a função objetivo a ser minimizada como

$$S_M(z_1, \dots, z_n) = \sum_{i \neq j} \left(d_{ij} - \|z_i - z_j\|\right)^2 \tag{3.1}$$

Essa função é conhecida como escalonamento por mínimos quadrados ou de Kruskal-Shephard. A ideia é encontrar uma representação em menor dimensionalidade que preserve os pares de distâncias da melhor maneira possível. A aproximação 3.1 é feita em termos de distância em vez de quadrados de distância, o que resulta em cálculos mais simples. Um algoritmo de descida gradiente é utilizado para minimizar S_M.

Uma variação do escalonamento por mínimos quadrados é o chamado mapeamento Sammon, que minimiza

$$S_{Sm}(z_1, \dots, z_n) = \sum_{i \neq j} \frac{\left(d_{ij} - \|z_i - z_j\|\right)^2}{d_{ij}} \tag{3.2}$$

Nesse caso mais ênfase é colocada na preservação de pares de distâncias menores (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Há na literatura mais dois métodos populares, o escalonamento multidimensional clássico (cMDS) e o não métrico (nMDS).

3.1 ESCALONAMENTO MULTIDIMENSIONAL CLÁSSICO (cMDS)

No escalonamento multidimensional clássico, ao contrário do que visto anteriormente, utilizamos as medidas de similaridade s_{ij}, usualmente dada pelo produto interno centrado $s_{ij} = \langle x_i - \bar{x}; x_j - \bar{x} \rangle$, porém, se tivermos dissimilaridades em vez de produtos internos, podemos convertê-las em produtos internos centrados - com mais eficácia se forem distâncias euclidianas. Portanto, temos que minimizar

$$S_C(z_1, \dots, z_n) = \sum_{i,j} \left(s_{ij} - \langle z_i - \bar{z}; z_j - \bar{z} \rangle\right)^2 \tag{3.3}$$

sobre $z_1, \dots, z_n \in \mathbb{R}^q$. cMDS é um método muito popular pelo fato de haver uma solução explícita em termos de autovetores.

Considere S como a matriz de produto interno centrado com elementos $\langle x_i - \bar{x}; x_j - \bar{x} \rangle$. Temos $\lambda_1 > \lambda_2 > \cdots > \lambda_q$ representando os q maiores autovalores de S, com autovetores associados $E_q = (e_1, e_2, \dots, e_q)$ e D_q representando a matriz diagonal com entradas $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_q}$. As soluções z_i para o escalonamento multidimensional clássico são dadas pelas linhas de $E_q D_q$ (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

3.2 ESCALONAMENTO MULTIDIMENSIONAL NÃO MÉTRICO (nMDS)

Os métodos dados acima são referidos como métodos métricos de escalonamento multidimensional, no sentido de que as dissimilaridades, ou similaridades, reais são aproximadas. Em contrapartida, o escalonamento não métrico de Shephard-Kruskal usa, efetivamente, apenas classificações (*ranks)* e visa minimizar

$$S_{NM}(z_1, \dots, z_n) = \frac{\sum_{i \neq j} \left[\|z_i - z_j\| - \theta(d_{ij})\right]^2}{\sum_{i \neq j} \|z_i - z_j\|^2} \tag{3.4}$$

sobre z_i e uma função crescente arbitrária θ. Fixando θ, z_i é minimizado por gradiente descendente, então z_i é fixado e o método da regressão isotônica é utilizado para encontrar a melhor aproximação monotônica $\theta(d_{ij})$ para $\|z_i - z_j\|$. Esses passos são iterados até que a solução se estabilize (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

3.3 QUALIDADE DO AJUSTE

3.3.1 Coeficiente de mensuração

A representação na dimensão reduzida q é dada pelos autovalores, λ_i, correspondentes aos q maiores autovetores. Para verificar a adequação da representação na dimensão q, quando utilizamos técnicas de escalonamento multidimensional clássico (ou outro método métrico), temos o seguinte coeficiente

$$P_q = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{3.5}$$

Valores de P_q acima de 0,80 indicam que as observações estão bem representadas na nova dimensão.

3.3.2 Stress

O *Stress* (KRUSKAL, 1964) indica o quanto a ordenação das distâncias na dimensão q está de acordo com a ordenação na dimensão original. Utilizada em escalonamento multidimensional não métrico, a medida *Stress* é definida como

$$Stress\ (q) = \left[\frac{\sum_{i<j}\left(d_{ij} - d_{ij}^{q}\right)^2}{\sum_{i<j}\left(d_{ij}\right)^2}\right]^{1/2}, \tag{3.6}$$

sendo d_{ij} as distâncias originais entre os objetos e d_{ij}^{q} as distâncias na nova dimensão.

Para verificar a qualidade do ajuste, Kruskal (1964) sugere o critério descrito na TABELA 1. Quanto maior o valor do *Stress*, pior será o ajuste.

Tabela 1 Critério Stress

Stress (%)	Qualidade do ajuste
10	Razoável
5	Bom
2,5	Excelente
0	Perfeito

É importante ressaltar que, por se tratar de aprendizagem não supervisionada, a qualidade do ajuste dada acima nem sempre acompanha a facilidade de interpretação dos dados, sendo necessário avaliar ambos em conjunto.

4 ANÁLISE DE AGRUPAMENTO

Como introduzido anteriormente, o objetivo da análise de agrupamento é alocar as observações em grupos (*clusters*) de modo que as dissimilaridades de pares entre aqueles atribuídos ao mesmo grupo tendam a ser menores do que aquelas em grupos diferentes, para isso é necessária a aplicação de um algoritmo de agrupamento.

Os algoritmos de agrupamento se enquadram, de forma geral, em métodos hierárquicos e métodos de particionamento (KASSAMBARA, 2017). É importante ressaltar que devido à grande heterogeneidade das aplicações de agrupamento, os métodos são normalmente desenvolvidos para determinadas classes de problemas, ou seja, não existe um método que seja genérico a ponto de obter bons resultados em todas as aplicações de agrupamento.

4.1 MÉTODOS DE PARTICIONAMENTO

Os métodos de particionamento, ou métodos não-hierárquicos, buscam encontrar a melhor partição dos *n* objetos em *k* grupos através da formação de uma partição inicial. Em geral, adota-se *k* observações como “sementes” do algoritmo para a formação do mesmo número de grupos.

Desses, os métodos mais utilizados são baseados em um ponto central (média dos atributos dos objetos – K-médias) ou em um objeto representativo para o grupo (K-medoides), para esse trabalho, consideraremos apenas o algoritmo K-médias, porém, K-medoides é construído de maneira análoga, com maior custo computacional.

4.1.1 Agrupamento por K-médias

K-médias (MACQUEEN, 1967) é um dos mais populares algoritmos de aprendizagem estatística não supervisionada, utilizado para particionar um determinado conjunto de dados em *k* grupos e destina-se a situações em que todas as variáveis são do tipo quantitativo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

O objetivo do algoritmo é minimizar, de forma iterativa, a distância dos objetos a um conjunto de *k* centros, também chamados de centroides, dado por $\chi = \{x_1, x_2, \dots, x_k\}$. A distância entre um ponto p_i e um conjunto de clusters $d(p_i, \chi)$ é definida como sendo a distância do ponto ao centroide mais próximo dele. Temos que a função a ser minimizada, de maneira geral, é dada por:

$$d(P, \chi) = \frac{1}{n} \sum_{i=1}^{n} d(p_i, \chi)^2 \tag{4.1}$$

O principal argumento é o parâmetro *k* (número de grupos e centroides) definido previamente pelo usuário. Isto costuma ser um problema, pois normalmente não se sabe quantos grupos existem a priori. Na SEÇÃO 5 serão apresentadas técnicas de estimação do número ideal de grupos a fim de contornar esse problema.

O algoritmo K-médias pode ser descrito da seguinte maneira:

Algoritmo 4.1 *Agrupamento por K-médias*

1. Escolher o número *k* de grupos a serem criados
2. Selecionar aleatoriamente k objetos do conjunto de dados para serem os centros inicias dos grupos
3. Associar cada observação ao centro mais próximo com base na medida de distância (ou algoritmo) definida
4. Recalcular o centro de cada grupo dado pela média dos objetos que o formam
5. Repetir os passos 3 e 4 até nenhum objeto mudar de grupo ou o máximo de iterações ser atingido (por padrão o software R utiliza 10 iterações como máximo)

Esse algoritmo iterativo tende a convergir para um mínimo da função 4.1 definida acima. Um eventual problema é a má separação dos conjuntos no caso de uma má inicialização dos centros, que é feita de forma arbitrária no início da execução.

Outra, e a principal, desvantagem que pode afetar a qualidade dos resultados é a escolha do número de grupos feita pelo usuário. Um valor pequeno de *k* pode causar a junção de dois grupos naturais, analogamente, um valor grande para *k* pode fazer com que um grupo natural seja quebrado artificialmente em dois ou mais.

A grande vantagem é que esse algoritmo é simples e extremamente veloz, geralmente convergindo em poucas iterações para uma configuração estável na qual nenhum objeto está designado a um grupo cujo centro não lhe seja o mais próximo.

4.1.1.1 Algoritmo de Hartigan-Wong

O algoritmo padrão e mais eficiente para o cálculo de K-médias é o de Hartigan-Wong (1979) (HASTIE; TIBSHIRANI; FRIEDMAN, 2009) que define a variância total dentro do grupo como a soma dos quadrados das distâncias euclidianas entre o objeto e seu respectivo centroide:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - u_k)^2, \tag{4.2}$$

em que x_i é uma observação pertencente ao cluster C_k e u_k representa o valor médio dos pontos atribuídos ao grupo C_k, ou seja, o valor do centroide.

Cada observação é atribuída a um grupo de modo que a soma dos quadrados da distância da observação ao centro do grupo atribuído seja minimizada.

A variância total dentro do grupo é definida como segue:

$$var.total = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - u_k)^2 \quad (4.3)$$

Esse valor mede a compactação (ou ajuste) do agrupamento, então o ideal é que seja o menor possível (KASSAMBARA, 2017).

4.1.2 Qualidade do agrupamento

Para responder sobre a qualidade do agrupamento particionado, é apresentada a estatística R^2 conforme determinada abaixo

$$R^2 = \frac{SSB}{SST}, \quad (4.4)$$

em que SSB é a soma dos quadrados total entre os grupos (*betweenss*) e *SST* é a soma dos quadrados total. Quanto mais próximo de 1, melhor a qualidade do agrupamento em questão. Essa medida também pode ser utilizada para comparar o agrupamento para diferentes valores k de grupos.

4.2 MÉTODOS HIERÁRQUICOS

Como visto em 4.1, os resultados ao aplicarmos os algoritmos de particionamento K-médias ou K-medoides dependem da escolha do número de grupos a ser encontrado como configuração inicial. Em contrapartida, os métodos hierárquicos de particionamento não precisam de tal especificação, em vez disso, eles exigem que o usuário especifique uma medida de dissimilaridade entre grupos (disjuntos) de observações, tais medidas são chamadas de métodos de ligação (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Como o nome sugere, esses métodos produzem representações hierárquicas nas quais os grupos em cada nível da hierarquia são criados recursivamente pela fusão de grupos no nível anterior. No nível mais baixo, cada grupo é representado por uma única observação, enquanto que no nível mais alto há apenas um grupo contendo todos os dados.

Há duas principais abordagem em agrupamento hierárquico, a abordagem aglomerativa (*bottom-up*) e a divisiva (*top-down*). A primeira, também chamada de algoritmo AGNES, opera de baixo para cima, ou seja, cada objeto é inicialmente considerado como um grupo e, a cada etapa do algoritmo, os dois grupos mais semelhantes são combinados em um novo grupo maior. Esse procedimento é iterado até que todas as observações sejam membras de um único grande grupo.

Por outro lado, a abordagem divisiva, ou DIANA, opera de cima para baixo, sendo uma ordem inversa da abordagem aglomerativa. Inicialmente todos os objetos são incluídos em um único grupo e, em cada etapa da iteração, o grupo mais heterogêneo é dividido em dois. O algoritmo segue sucessivamente até que todos os objetos estejam em seus próprios grupos.

A abordagem divisiva não foi estudada tão extensivamente quanto os métodos de aglomeração na literatura de análise de agrupamento. Foi mais fortemente explorada na literatura de engenharia (GERSHO; GRAY, 1992) no contexto de compressão, porém, uma vantagem dos métodos divisivos sobre os aglomerativos pode ocorrer quando o interesse é o particionamento dos dados em um número pequeno de grupos, pois essa abordagem trabalha melhor com grupos largos, ou seja, compostos por muitas observações (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A estrutura hierárquica formada pela união entre os elementos é comumente representada através de um dendrograma, que mostra intuitivamente a ordem do agrupamento. Em termos da abordagem aglomerativa, quanto mais alta a linha ligando dois grupos, mais tarde foi feito seu agrupamento, o contrário pode ser dito para a abordagem divisiva. Logo, essa altura ligando dois grupos é proporcional à sua distância.

4.2.1 Métodos de ligação

Como introduzido anteriormente os métodos de agrupamento hierárquicos exigem que o usuário especifique uma medida de dissimilaridade entre grupos, chamada de método de ligação ou aglomeração de grupos.

Focaremos em algoritmos aglomerativos, visto que é mais fácil unir grupos do que separá-los, além do que foi introduzido em 4.2. Neles, o método de ligação será responsável na fusão dos grupos dois as dois, inicialmente compostos por apenas um objeto, a fim de minimizar a variância interna do grupo e maximizar essa variância entre um grupo e outro sucessivamente até que haja apenas um grupo (KAUFMAN; ROUSSEEUW, 1990).

Sejam G e H dois grupos distintos, a dissimilaridade $d(G,H)$ é calculada a partir do conjunto de dissimilaridades de observações d_{ij} no qual um membro do par (i) está em G e outro (j) em H (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Há diferentes métodos de ligação, ou seja, diferentes formas de calcular $d(G,H)$. Hastie et al. (2009) consideram as ligações única, completa e média como as mais popularmente efetivas. Além dessas, consideraremos o método de variação mínima de Ward.

4.2.1.1 Ligação única ou mínima

Também chamada de técnica do vizinho mais próximo, a ligação única (SL) calcula todas as dissimilaridades entre os elementos nos grupos G e H e utiliza a menor delas como critério de ligação.

$$d_{SL}(G,H) = \min\left(d_{ij}\right), i \in G, j \in H \tag{4.5}$$

Uma preocupação de usar esse método é que as vezes pode produzir encadeamento entre grupos, ou seja, vários grupos podem ser unidos porque um de seus objetos está muito próximo do objeto de um grupo separado. Este problema é específico da ligação simples, visto que a menor distância entre os pares é o único valor levado em consideração.

4.2.1.2 Ligação completa ou máxima

De maneira contrária a ligação única, a ligação completa (técnica do vizinho mais longe) considera como critério de ligação a maior dissimilaridade entre os pares de observações. É um método que assegura que todos os itens de um cluster estão a uma distância mínima um do outro e tende a produzir grupos mais compactos.

$$d_{CL}(G,H) = \max\left(d_{ij}\right), i \in G, j \in H \tag{4.6}$$

Embora a ligação completa resolva o problema do encadeamento visto anteriormente, os casos discrepantes influenciam fortemente na medida do vizinho mais distante, exacerbando os efeitos dos valores extremos possivelmente impedindo que os grupos próximos se fundam.

4.2.1.3 Ligação média

O método da ligação média (GA) utiliza a dissimilaridade média entre pares de objetos como critério para a elaboração da matriz de distâncias:

$$d_{GA}(G,H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{j \in H} d_{ij}, \tag{4.7}$$

na qual N_G e N_H são os respectivos números de observações em cada grupo.

A ligação média considera todos os membros dos grupos cuja distância está sendo calculada. Consequentemente, esse método é menos influenciado por valores discrepantes – o que ocorre com as ligações única e completa.

4.2.1.4 Método de variação mínima de Ward

O método de variação mínima, ou mínima variância, de Ward (WARD, 1963) é direcionado na criação de grupos de igual tamanho, para isso, parte da soma dos quadrados dos erros (SQE) de cada grupo – soma dos quadrados dos desvios para o centroide do grupo. Somam-se os SQEs de todos os grupos, gerando assim o SQET.

Com agrupamento hierárquico, a SQE começa em zero, visto que cada observação está em seu próprio grupo, e então cresce à medida que juntamos os grupos. A intenção é detectar qual dessas uniões produz o menor aumento de SQE, buscando assim, maximizar a verossimilhança em cada nível de hierarquia.

$$\Delta(G,H) = \frac{N_G N_H}{N_G + N_H} \|\overrightarrow{m_G} - \overrightarrow{m_H}\|^2, \tag{4.8}$$

na qual Δ é chamado de custo de fusão da combinação dos clusters G e H, $\overrightarrow{m_G}$ e $\overrightarrow{m_H}$ são os respectivos centros dos grupos.

De maneira geral, o método de mínima variância de Ward tende a formar grupos com número pequeno de observações, na maioria das vezes com mesmo formato e número de observações. Assim como os dois primeiros métodos, é também muito sensível a valores extremos.

4.2.2 Avaliação da estrutura hierárquica

Para avaliarmos a eficácia da formação dos grupos em uma estrutura de agrupamento hierárquico, temos, em AGNES, o coeficiente aglomerativo (ou coeficiente de aglomeração) e de maneira análoga em DIANA, o coeficiente divisivo ou de divisão. Kaufman e Rousseeuw (1990) definem que os coeficientes descrevem a força da estrutura de agrupamento que foi obtida pelo método de ligação utilizado.

Ambos coeficientes são na verdade a média das alturas (ou comprimentos) normalizadas em que os grupos são formados, ou seja, as alturas que são dados pelo dendrograma.

O coeficiente de aglomeração e de divisão variam de 0 a 1. Valores mais próximos do máximo sugerem uma estrutura de agrupamento mais equilibrada, enquanto valores mais próximos de 0 indicam que os grupos foram menos bem formados. No entanto tendem a se tornar maior à medida que o tamanho do conjunto de dados aumenta, portanto, o uso não é recomendado para comparar conjuntos de dados de tamanhos muito diferentes (KAUFMAN; ROUSSEEUW, 1990).

5 DETERMINAÇÃO DO NÚMERO IDEAL DE GRUPOS

A grande parte dos algoritmos de agrupamento presentes na literatura requer que o usuário escolha previamente o número *k* de clusters que, juntamente com a medida de distância, é crucial para a eficácia do algoritmo e pode mudar completamente as conclusões da análise (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; KAUFMAN; ROUSSEEUW, 1990).

Sarle (1983) considera estimar o número de grupos presentes nos dados como o problema mais difícil na análise de agrupamento. Nesse contexto, a pergunta que desejamos responder é qual valor de k deve ser escolhido quando não se tem conhecimento a priori sobre o número de grupos em determinado conjunto de dados.

Há uma série de critérios para a determinação do número ideal de grupos, dentre eles, os métodos cotovelo, silhueta média, estatística de lacuna e CCC são os mais populares (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; SARLE, 1983).

5.1 MÉTODO COTOVELO

Como introduzido, métodos de agrupamento buscam alocar as informações em grupos tal que a variação total dentro dos mesmos, também conhecida como soma de quadrados dentro do grupo (5.1), é minimizada (KAUFMAN; ROUSSEEUW, 1990).

O método cotovelo (*Elbow method*) consiste em olhar a porcentagem de variância explicada como uma função do número de clusters. Esse método existe sobre a ideia de que se deve escolher um número de grupos *k* de forma que *k* + 1 não forneça uma modelagem muito melhor dos dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

$$wss = \ min\left(\sum_{k=1}^{k} W(C_k)\ \right), \tag{5.1}$$

na qual C_k é o k-ésimo cluster e $W(C_k)$ é a variação dentro do cluster. A soma total de quadrados dentro do grupo (*wss*) mede a compactação do agrupamento e, por isso, desejamos que seja o menor possível.

Assim, o algoritmo 5.1 é utilizado para definir o número ideal de grupos de acordo com esse método.

Algoritmo 5.1 *Método cotovelo*

1. Utilizar o algoritmo de agrupamento em questão para valores diferentes de *k*. Por exemplo, variando *k* de 1 a 10 grupos.
2. Para cada valor de k, calcular a soma de quadrados total dentro do grupo (*wss*).
3. Traçar a curva de *wss* em função do número de clusters *k*.

A localização de uma dobra (cotovelo) no gráfico é considerada como um indicador do número ideal de grupos, ou seja, onde a adição de outro grupo não muda muito a configuração da variabilidade explicada.

5.2 MÉTODO DE SILHUETA MÉDIA

A abordagem da silhueta média (average silhouette method) mede a qualidade de um agrupamento, ou seja, determina o quão bem cada objeto se encontra em seu grupo, levando em consideração as diferenças entre as distâncias que um objeto tem de outros no mesmo grupo e a distância que ele tem de outros objetos em grupos diferentes (KAUFMAN; ROUSSEEUW, 1990).

Para esse método é utilizada a largura da silhueta que, para um objeto *i* no grupo C_i, é definida através da equação abaixo:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \tag{5.2}$$

na qual $b_i = \min\left(\sum_j d(i,j)/|C_j|\right)$ com $C_i \neq C_j$ e $a_i = \min\left(\sum_j d(i,j)/|C_i|\right)$ com $C_i = C_j$.

Kaufman e Rousseeuw (1990) consideram a largura da silhueta como um indicador de performance variando de -1 a 1. Um alto valor de s_i indica que os objetos foram bem agrupados, além disso, adicionam o seguinte sumário com as definições:

- A média de s_i para determinado objeto i em um grupo é chamada silhueta média do cluster.

- A média de s_i para todos os objetos é chamada de silhueta média para todo conjunto de dados e é denotada como $\bar{s}(k)$, sendo k o número de grupos.

Com isso podemos afirmar que o número ideal de grupos, dado o método da silhueta média, é aquele que maximiza $\bar{s}(k)$ dentro dos possíveis valores de *k* (KAUFMAN; ROUSSEEUW, 1990).

5.3 CUBIC CLUSTERING CRITERION (CCC)

Um indicador bastante popular é o critério CCC (*cubic clustering criterion*), de Sarle (1983), que, em algoritmos de agrupamento testa a hipótese H_0 de que os dados foram amostrados de uma distribuição uniforme, contra a hipótese H_1 de que os dados foram amostrados de uma mistura de distribuições normais multivariadas esféricas com variâncias e probabilidades amostrais iguais.

Sarle (1983), em seu método, obteve uma aproximação para a distribuição do coeficiente de determinação R^2, neste caso, entendido como a proporção de variância com a qual cada grupo contribui para a variância total dos dados (soma de quadrados total), sob a hipótese nula. O CCC é então obtido ao se comparar o R^2 observado nos dados após o agrupamento com seu valor esperado sob hipótese nula, utilizando uma transformação para estabilização da variância aproximada. A expressão para esse critério é dada abaixo

$$CCC = \log\left[\frac{1 - E(R^2)}{1 - R^2}\right].v, \tag{5.3}$$

na qual v é o número de variáveis na base de dados.

Valores positivos de 5.3 significam que o coeficiente R^2 dos dados é maior do que se esperaria de uma amostra com distribuição uniforme (rejeição de H_0), indicando assim a presença de grupos nos dados.

A melhor maneira de visualizar o CCC é plotando seu valor em função do número k de grupos, picos nos quais o valor de 5.3 excede 2 ou 3 mostram um bom agrupamento, ou seja, uma boa estimativa para o número ideal de grupos (sendo preferível o maior valor possível).

Uma possível vantagem para esse método se dá quando o conjunto de dados a ser analisado é proveniente de uma mistura de distribuições normais multivariadas esféricas com variâncias e probabilidades amostrais iguais. Entretanto, caso os grupos sejam muito pequenos (menos de 10 observações) o critério CCC pode ser falho (SARLE, 1983).

5.4 MÉTODO DE ESTATÍSTICA DE LACUNA

Publicado por Tibishirani, Walther e Hastie (2001), o método de estatística de lacuna (Gap statistic) pode ser aplicado em qualquer método de agrupamento a fim de indicar o número *k* ideal de grupos. A estatística de lacuna compara a variação total dentro dos grupos para diferentes valores de *k* com seus respectivos valores esperados sob distribuição de referência nula dos dados, ou seja, uma distribuição sem agrupamento óbvio.

O conjunto de dados de referência é gerado através de simulações de Monte Carlo do processo de amostragem, isto é, para cada variável (x_i) no conjunto de dados, calculamos seu intervalo $[\min(x_i), \max(x_j)]$ e geramos valores para os *n* pontos através de uma distribuição uniforme dentro desse intervalo.

Para os dados observados e os dados de referência, o total da variação dentro dos grupos é computado com base em diferentes valores de *k*. Enfim, a estatística de lacuna para determinado *k* é definida como segue (TIBSHIRANI; WALTHER; HASTIE, 2001):

$$Gap_n(k) = E_n^* \log(W_k) - \log(W_k)\,, \tag{5.4}$$

na qual E_n^* denota a expectativa sob um tamanho de amostra n da distribuição de referência e W_k, como visto anteriormente, é a variação (dispersão) dentro do grupo.

A estimativa do número ideal de grupos $(\hat{k})$ é dada pelo valor que maximiza 5.4. Podemos notar que a estimativa é bastante generalizada, ou seja, além de ser aplicável para qualquer método de agrupamento, pode ser utilizada independente da métrica de distância. Hastie et al. (2009) afirmam que a medida de distância é extremamente eficaz e pode ser utilizada sozinha na determinação do número ideal de grupos em um conjunto de dados.

6 APLICAÇÕES

Neste capítulo serão realizadas aplicações em dois conjuntos de dados para os métodos de agrupamento e redução de dimensionalidade apresentados. Toda a análise será feita no software R (R CORE TEAM, 2020) e os códigos estão apresentados no apêndice desde trabalho.

6.1 IRIS

Para ilustrar os métodos de agrupamento, tal como a técnica de escalonamento multidimensional, primeiramente usaremos o conjunto de dados disponível no próprio software R – íris (FISHER, 1936). Esse banco de dados fornece as medidas, em centímetros, das variáveis comprimento e largura tanto das pétalas quanto das sépalas, respectivamente, para 50 flores de cada uma das 3 espécies de íris.

6.1.1 Redução de dimensionalidade

Por se tratar de um conjunto de dados métrico, para a redução de dimensionalidade foi realizado o escalonamento multidimensional clássico e, como introduzido anteriormente, a medida de dissimilaridade euclidiana é a que mais se adapta nesse caso, portanto, temos que a matriz de distâncias será calculada conforme 2.3.

A FIGURA 1 apresenta o resultado do cMDS (utilizando a função *cmdscale*) com uma dimensão para o conjunto de dados em questão, enquanto a FIGURA 2 apresenta o resultado para duas dimensões. Em ambos os casos, podemos identificar principalmente um agrupamento dado mais à esquerda do gráfico, devido sua heterogeneidade com o restante dos dados. Através das figuras podemos claramente ver dois agrupamentos de pontos, indicando que $k \geq 2$.

Através do cálculo do coeficiente de mensuração obtivemos $P_1 = 0{,}925$ e $P_2 = 0{,}978$, ou seja, a representação em uma dimensão para o conjunto original dos dados foi muito bem adequada e será utilizada na visualização do método de agrupamento por K-médias.

Em termos de interpretabilidade, podemos dizer que a redução de quatro dimensões para uma possivelmente faz com que as características das flores sejam reduzidas a uma espécie de volume em $\mathbb{R}^4$, já em duas dimensões, podemos imaginar as áreas das pétalas e sépalas.

6.1.2 Análise de agrupamento

Com a redução de dimensionalidade, podemos utilizar o algoritmo de agrupamento K-médias para alocar os objetos em grupos distintos visando maximizar tanto a homogeneidade dentro dos grupos quanto a heterogeneidade entre grupos. Usaremos também métodos de agrupamento hierárquicos, porém, esses não consideram a redução de dimensionalidade e sim a hierarquia entre objetos.

Temos, para o método de particionamento, na TABELA 2, os valores do coeficiente R^2 para as medidas de dissimilaridade propostas neste trabalho (com

exceção da Mahalanobis por sua condição). O coeficiente será utilizado como critério de escolha da medida de distância a ser utilizada. Para o cálculo dessas medidas, os dados foram padronizados.

Figura 1 cMDS com uma dimensão

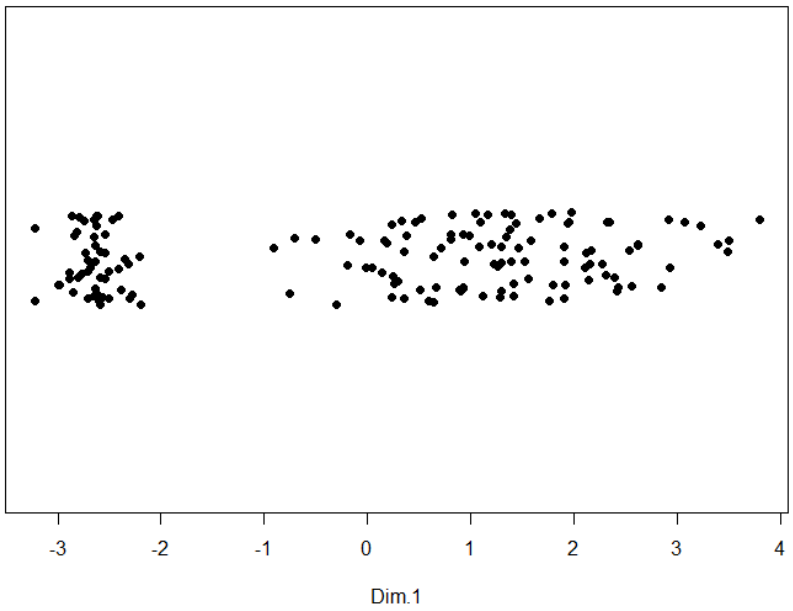


Figura 2 cMDS com duas dimensões

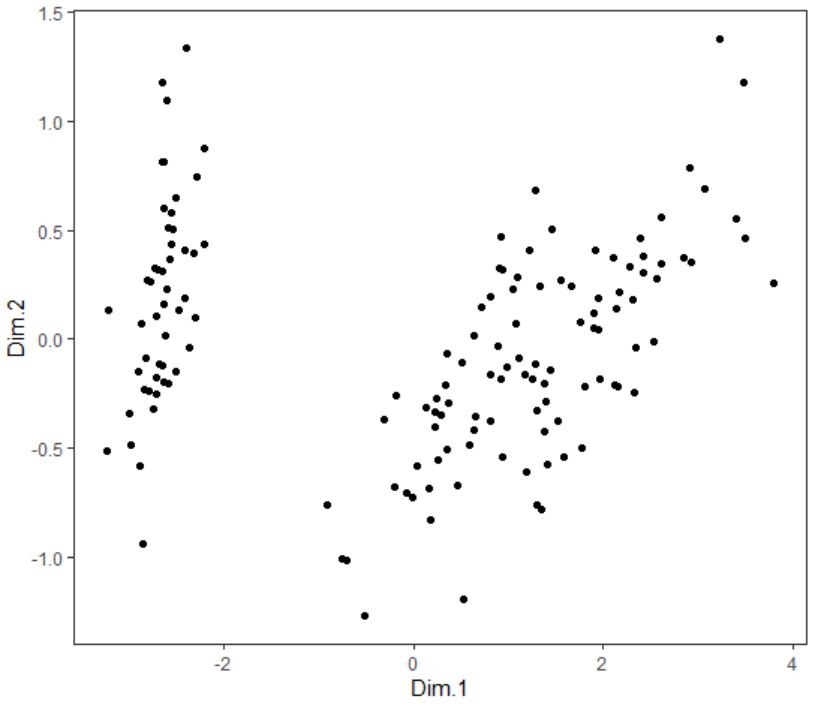


Tabela 2 Coeficiente R^2 para diferentes medidas de dissimilaridade

Dissimilaridade	R^2 $(k = 2)$	R^2 $(k = 3)$
Euclidiana	0,711	0,823
Manhattan	0,743	0,852
Gower	0,775	0,889

A escolha do número de grupos foi realizada através dos métodos descritos em seções anteriores. Pela FIGURA 3 podemos notar que o método da silhueta média indica que $k = 2$ é a melhor estimativa, enquanto que o método da estatística de lacuna indica que k deve ser 3 (selecionando através do valor máximo – linha azul escuro). Na FIGURA 4, através da soma de quadrados dentro do grupo (método do cotovelo) podemos escolher o valor $k = 3$, porém, vemos que $k = 2$ não seria uma má escolha. Por fim, temos o critério CCC, que nos fornece $k = 3$ como número ideal de grupos.

Portanto, considerando também a TABELA 2, computaremos o algoritmo K-médias utilizando distância de Gower e 3 centroides, ou seja, visando agrupar o conjunto de dados em 3 grupos. O resultado do agrupamento por K-médias (com redução de dimensionalidade dada pelo cMDS) é dado nas FIGURAS 5 e 6, nelas conseguimos ver a divisão dos três grupos. Em uma dimensão - FIGURA 5 - há uma leve sobreposição de dois grupos, enquanto que em duas dimensões - FIGURA 6 - não a vemos, indicando heterogeneidade entre os grupos, ou seja, mostrando uma boa alocação dos centroides que resulta em um bom agrupamento.

Figura 3 Métodos silhueta média e estatística de lacuna para escolha do número de grupos

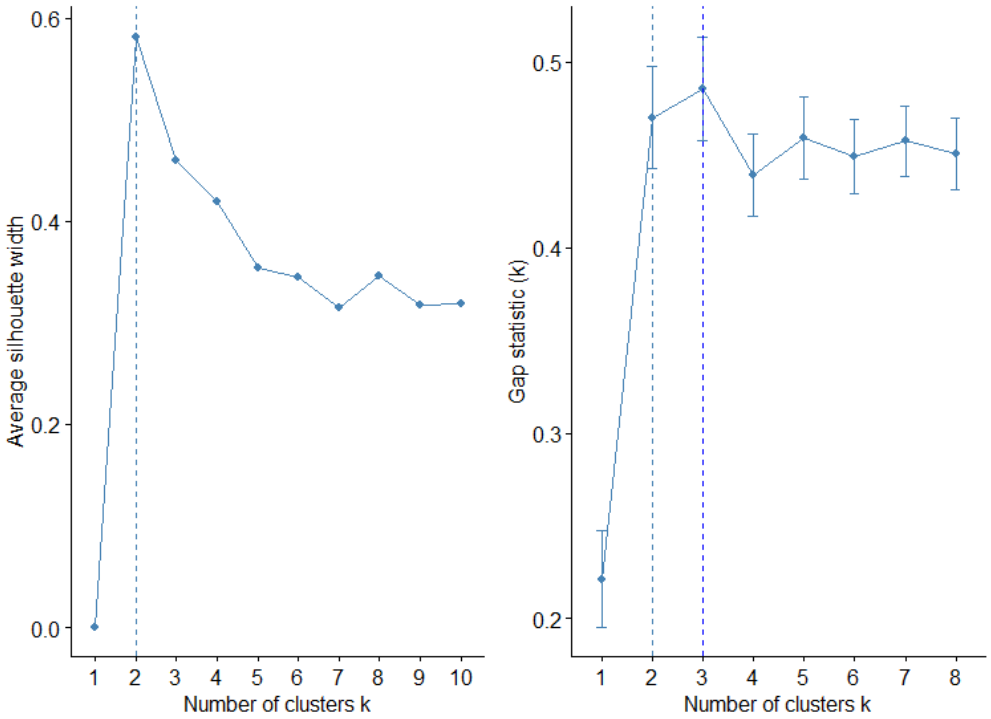
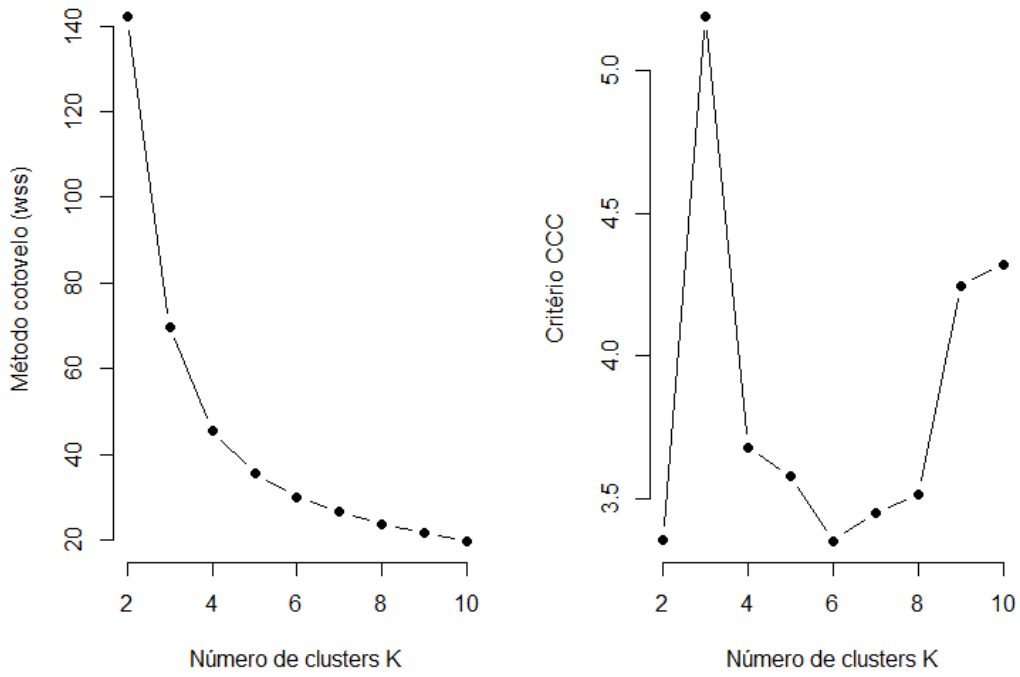


Figura 4 Métodos cotovelo (wss) e CCC para escolha do número de grupos



Os grupos são compostos por 50, 62 e 38 objetos, respectivamente da esquerda para a direita nas FIGURAS 5 e 6. Como vimos nas FIGURAS 1 e 2, dois grupos ficam muito próximos, causando uma deformação nesses grupos pelo método de particionamento K-médias. Apenas o grupo relativo à espécie setosa se mostrou composto pelas 50 observações dadas originalmente em iris.

A TABELA 3 apresenta os valores médios das variáveis para cada grupo obtido pelo método de particionamento utilizado. Com ela podemos ver mais claramente a diferença do primeiro grupo para os outros dois, cujos centroides ficam mais próximos um do outro, principalmente na variável largura da pétala.

Tabela 3 Resumo das variáveis originais por cada grupo dado por K-médias

Grupo	*Nº íris*	*Comprimento sépala*	*Largura sépala*	*Comprimento pétala*	*Largura pétala*
1	50	5,01	3,43	1,46	0,25
2	62	5,92	2,75	4,40	1,41
3	38	6,83	3,06	5,73	2,11

Figura 5 Agrupamento K-médias utilizando cMDS com uma dimensão

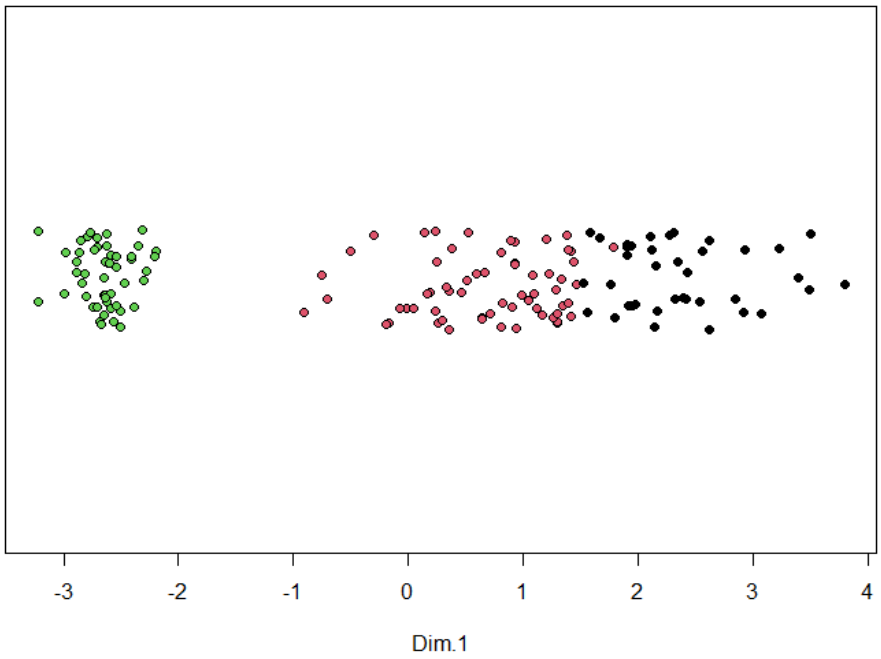
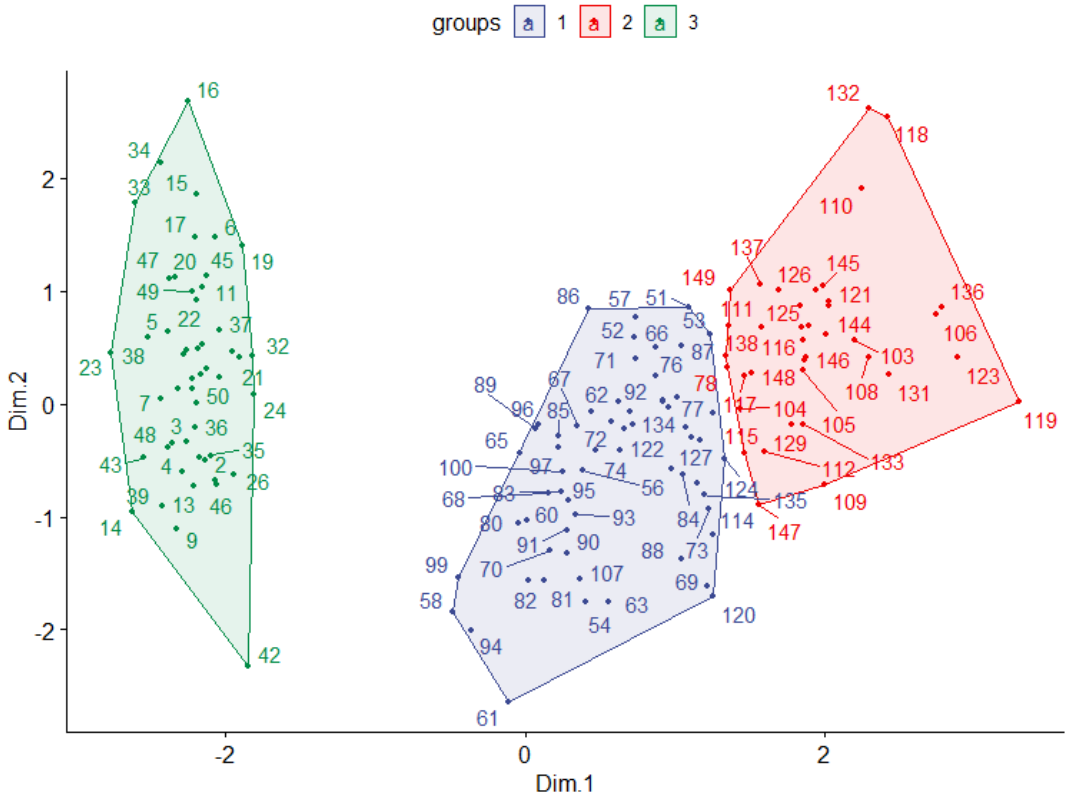


Figura 6 Agrupamento K-médias utilizando cMDS com duas dimensões



Em termos de agrupamento hierárquico, temos que a matriz de dissimilaridade será a mesma (dada pela distância de Gower), porém, além da medida de distância entre observações, temos que escolher a medida de distância entre os grupos - medida de ligação.

Comparando os valores do coeficiente aglomerativo, distribuídos na TABELA 4, para os métodos de ligação apresentados neste trabalho, temos que o método de variação mínima de Ward apresentou o maior coeficiente de aglomeração, seguido pelos métodos de ligação completa, média e única, respectivamente.

Analogamente ao método K-médias, o número ideal de grupos no conjunto de dados é $k = 3$, portanto, temos na FIGURA 7, o dendrograma representativo do agrupamento hierárquico aglomerativo do conjunto de dados iris e, similarmente, temos o dendrograma do algoritmo DIANA para o mesmo conjunto de dados na FIGURA 8.

Em AGNES, obtivemos 50, 65 e 35 observações, respectivamente, nos grupos formados, novamente um grupo criado conforme a alocação original do conjunto de dados e outros dois misturados entre si, desses, 3 objetos foram alocados em grupos diferentes do que visto com o método K-médias.

Com o algoritmo divisivo (DIANA), temos que o número de observações em cada grupo, respectivamente, foi de 54, 59 e 31, sendo o método menos eficaz para o agrupamento, quando comparamos com o conjunto original de dados.

Tabela 4 Coeficiente aglomerativo para quatro métodos de ligação

Ligação	*Coeficiente Aglomerativo*
Completa	0,958
Média	0,927
Única	0,855
Ward	0,990

Figura 7 Método hierárquico aglomerativo com variação mínima de Ward

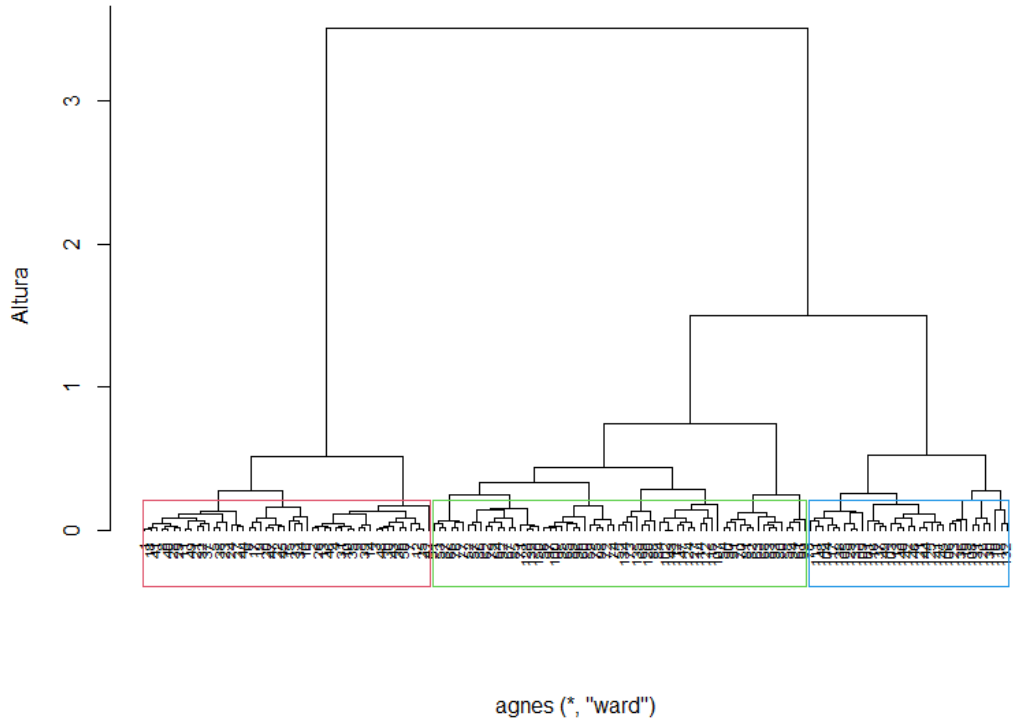
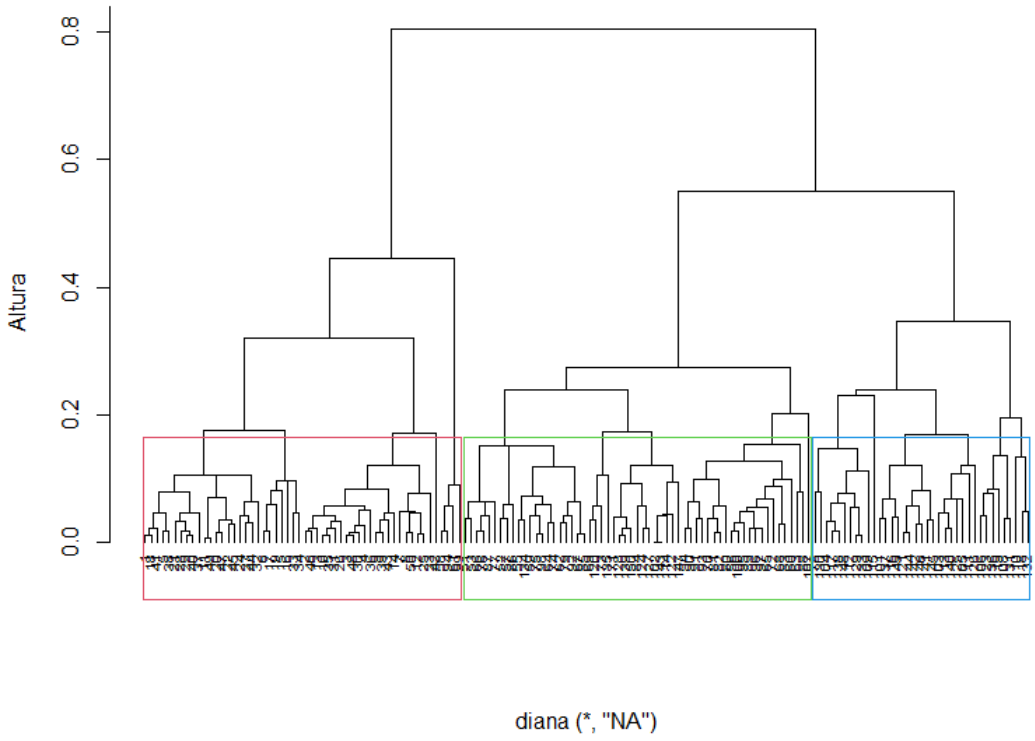


Figura 8 Método hierárquico divisivo



6.2 SEGMENTAÇÃO DE CLIENTES

O segundo conjunto de dados que utilizaremos diz respeito a um grande shopping que guarda informações de seus 200 clientes que assinam um cartão de filiação. Essas informações são dadas pelas variáveis sexo (0 = masculino; 1 = feminino), idade e renda anual (em milhar). Esse cartão é utilizado para fazer todas as compras no shopping, de forma que, através do histórico, é calculada a quarta variável: escore (ou pontuação) de gasto. Com essas informações, desejamos segmentar os clientes membros do shopping em grupos distintos.

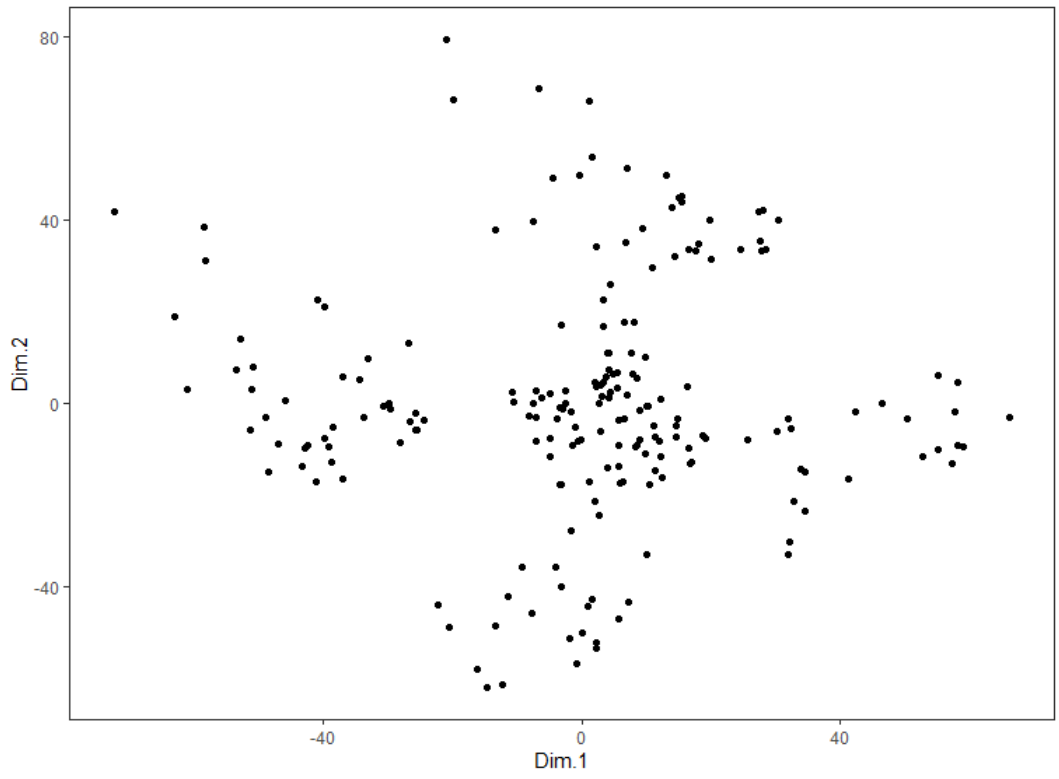
6.2.1 Redução de dimensionalidade

Ao contrário do que vimos anteriormente, o conjunto de dados em questão é híbrido com três variáveis métricas e uma variável não métrica. Portanto, aplicaremos o escalonamento multidimensional não métrico de Shephard-Kruskal (nMDS) utilizando a função *isoMDS* do pacote *MASS*.

A FIGURA 9 apresenta o resultado do nMDS em duas dimensões, que nos permite visualizar cinco aglomerações de pontos em um espaço bidimensional, ou seja, nos indica que o número de grupos no conjunto de dados $k \geq 5$.

O *Stress* (3.3.2) do nMDS bidimensional nos indica que a redução de dimensionalidade foi realizada de maneira razoável, porém, por questões de interpretabilidade e como temos apenas quatro variáveis no conjunto de dados iremos manter a representação em duas dimensões.

Figura 9 nMDS com duas dimensões



6.2.2 Análise de agrupamento

Por estarmos analisando um conjunto de dados misto (híbrido) as técnicas de agrupamento por particionamento não poderão ser utilizadas, visto que só podem ser aplicadas em conjuntos compostos inteiramente por variáveis quantitativas. Portanto, utilizaremos apenas os procedimentos AGNES e DIANA, que agrupam de maneira hierárquica.

Outra limitação que temos por trabalhar com o tipo híbrido de dados é a escolha da medida de dissimilaridade, visto que, no geral, há uma gama menor de opções, no caso deste trabalho, utilizaremos a distância de Gower.

Após a determinação da medida de dissimilaridade adotada, temos os determinação do número de grupos no conjunto de dados. As FIGURAS 10 e 11 apresentam os métodos cotovelo (wss), silhueta média, estatística de lacuna (GAP) e o critério CCC para determinação do número ideal de grupos. Todos os critérios apontam que $k = 6$ é a melhor estimativa para o verdadeiro número de grupos no conjunto de dados.

Figura 10 Métodos cotovelo (wss), silhueta média e estatística de lacuna para determinação do número de grupos

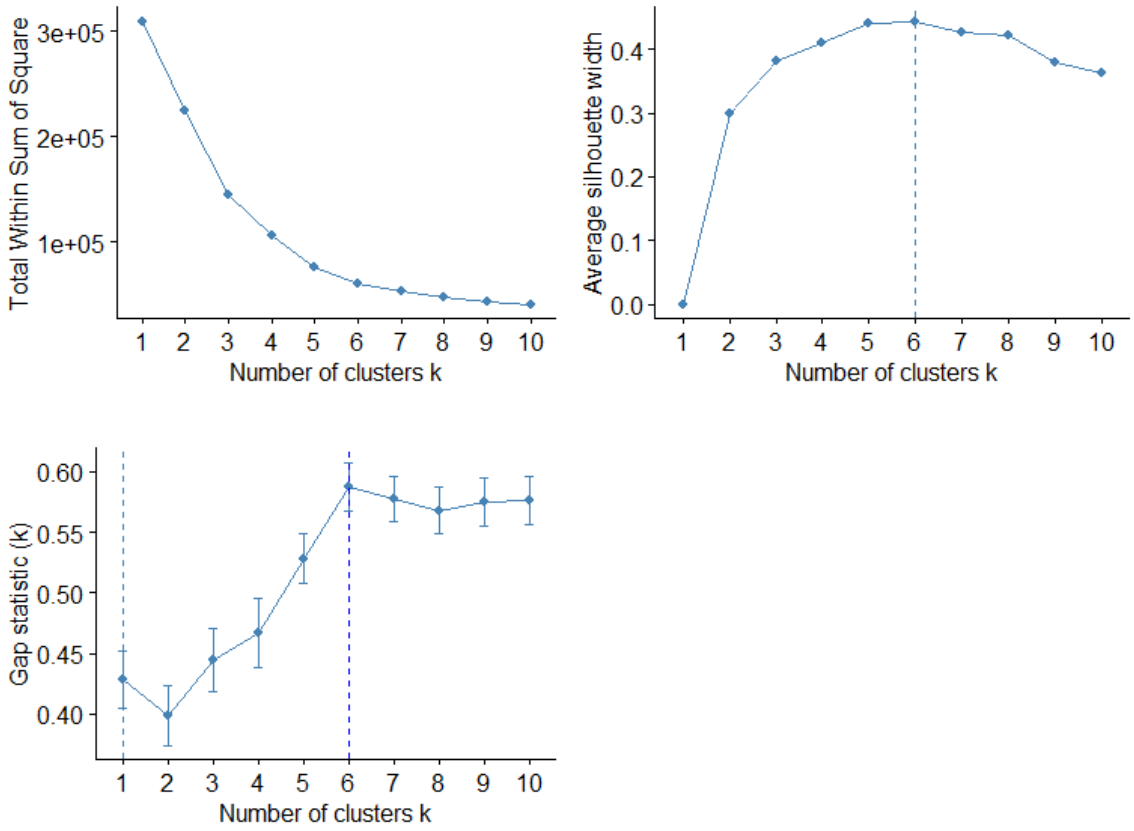
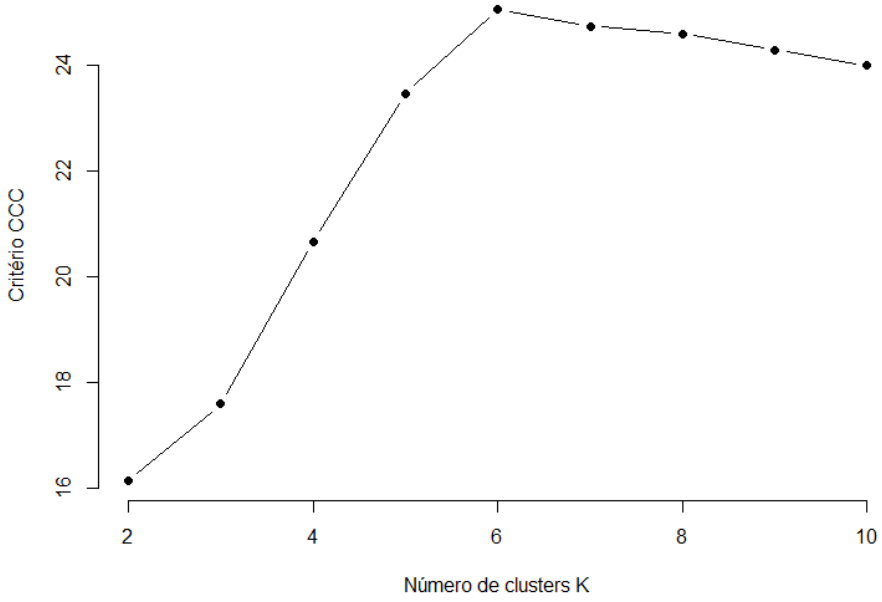


Figura 11 Critério CCC para determinação do número de grupos



Por fim, para os procedimentos hierárquicos, temos a escolha do método de ligação. A TABELA 5 apresenta os valores do coeficiente aglomerativo para os métodos de ligação propostos. Temos que o método de variação mínima de Ward apresentou o maior valor para o coeficiente, indicando uma melhor estrutura de aglomeração, seguido pelos métodos de ligação completa, média e única, respectivamente.

Tabela 5 Coeficiente aglomerativo para diferentes métodos de ligação

Ligação	*Coeficiente Aglomerativo*
Completa	0,956
Média	0,926
Única	0,880
Ward	0,991

Assim, temos na FIGURA 12, o dendrograma representativo do agrupamento hierárquico aglomerativo do conjunto de dados em questão. Obtivemos, nesse caso, 25, 18, 25, 20, 58 e 54 observações, respectivamente (da esquerda para a direita no dendrograma), nos grupos formados.

A TABELA 6 fornece informações a respeito das médias das variáveis métricas em cada grupo, além do valor da variável não métrica (cada grupo é composto 100% pelo valor de sexo indicado sendo 0 = masculino e 1 = feminino). Com esses dados, podemos concluir que, em termos de marketing, o shopping em questão deveria focar em atrair os clientes alocados no quarto grupo, pois são aqueles com alta renda e baixo escore de gastos.

Podemos observar também, através da TABELA 6 e da FIGURA 12, que, onde o conjunto se divide em apenas dois grupos, aproximadamente na altura 1,5, a divisão é dada por sexo, sendo o masculino no grupo 1 e o feminino no grupo 2, porém, nesse caso, perderíamos informações a respeito da renda e escore de gastos.

Figura 12 AGNES com variação mínima de Ward

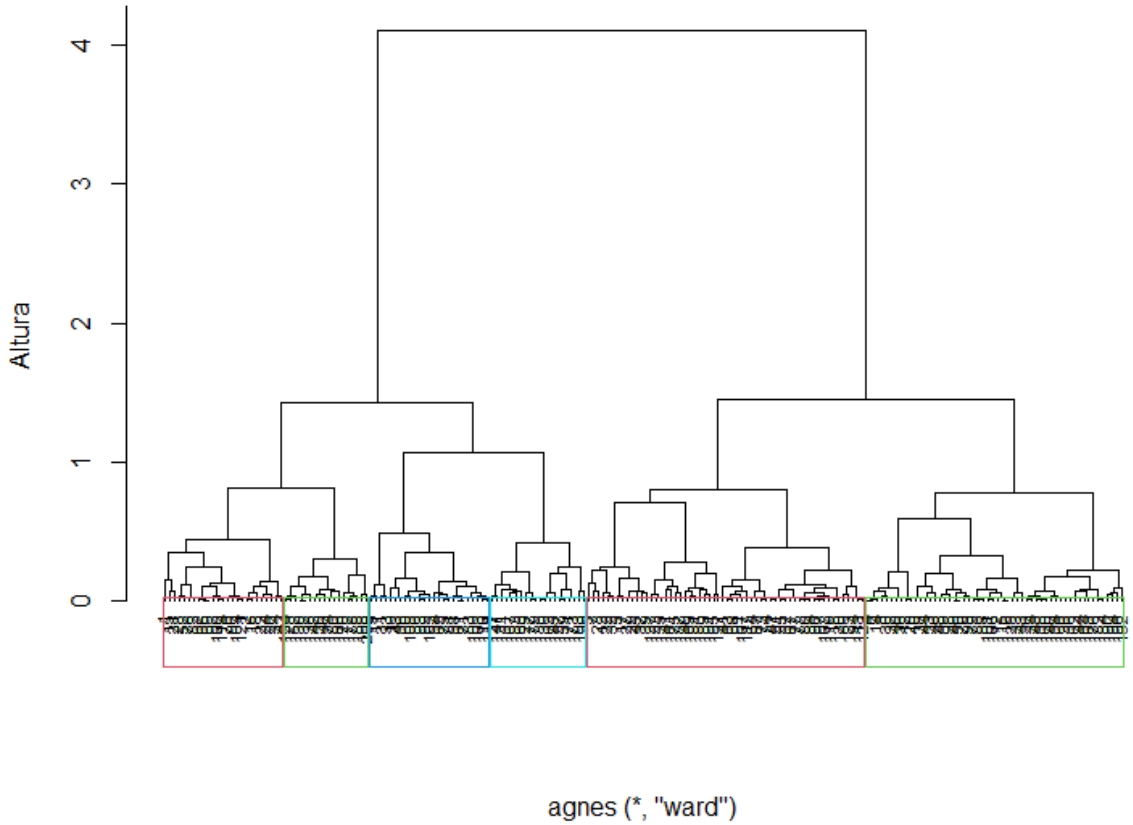


Tabela 6 Resumo das variáveis originais em cada grupo dado por AGNES

Grupo	*Nº clientes*	*Sexo*	*Idade*	*Renda (em milhar)*	*Escore de gasto*
1	25	0	25,7	40,4	59,0
2	18	0	33,3	87,1	82,7
3	25	0	58,8	47,8	41,0
4	20	0	39,5	85,2	14,0
5	58	1	47,4	58,3	35,7
6	54	1	28,1	60,3	68,5

Já com o algoritmo divisivo (DIANA), FIGURA 13, temos que o número de observações em cada grupo, respectivamente, foi de 39, 22, 27, 13, 42 e 57, nesse caso, o coeficiente de divisão dado foi de 0,945, indicando uma boa estrutura divisiva.

Analogamente ao que foi visto em AGNES, a TABELA 7 fornece informações a respeito das médias das variáveis métricas por grupo, além do valor da variável não métrica. Podemos observar que, nesse caso, as rendas anuais estão mais bem distribuídas pelos grupos, além de termos o mesmo número de grupos compostos por homens e mulheres, que se juntam aproximadamente na altura 0,6.

Em termos de marketing, o shopping deveria focar em atrair os clientes do grupo 3, que é composto por 27 clientes, em maioria do sexo masculino, que têm renda anual elevada, porém baixo escore de gasto.

Figura 13 DIANA

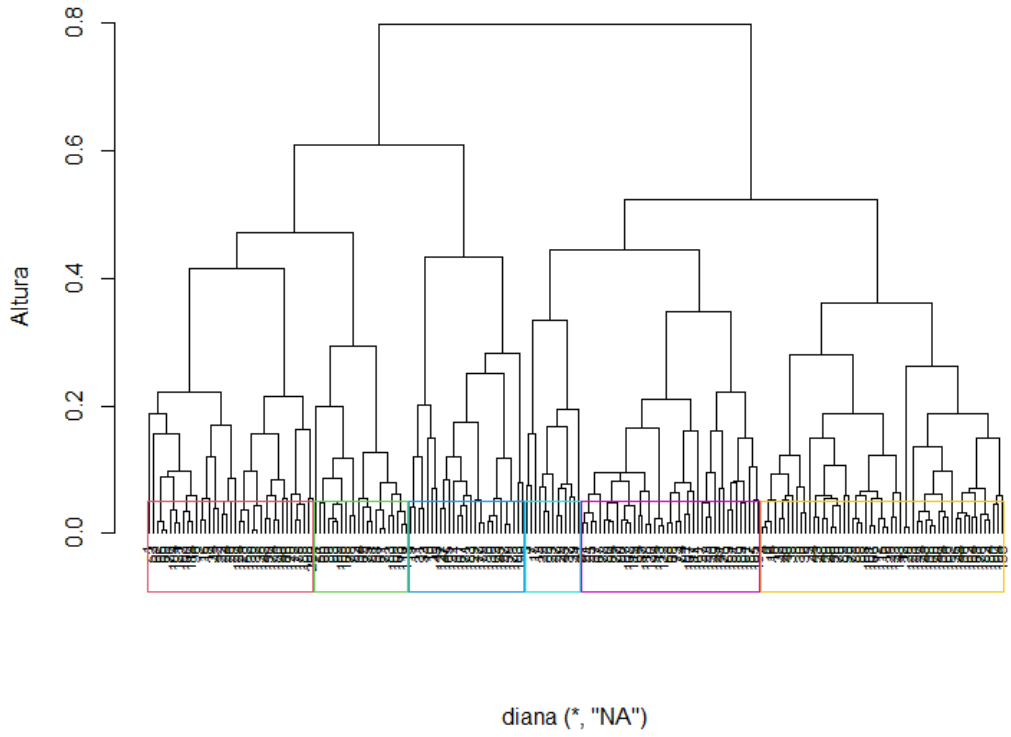


Tabela 7 Resumo das variáveis originais em cada grupo dado por DIANA

Grupo	*Nº clientes*	*Sexo*	*Idade*	*Renda (em milhar)*	*Escore de gasto*
1	39	0	28,0	62,2	72,1
2	22	0	56,4	52,9	48,8
3	27	0	43,4	69,9	14,2
4	13	1	44,2	28,2	20,3
5	42	1	49,3	68,3	39,3
6	57	1	28,4	59,7	67,7

7 CONSIDERAÇÕES FINAIS

O objetivo deste trabalho era explorar ao máximo os procedimentos de aprendizagem estatística não supervisionada dados pelas técnicas de agrupamento e escalonamento multidimensional, este para redução de dimensionalidade. Depois da revisão de conceitos, fizemos aplicações em dois conjuntos de dados comuns nessa área de estudo, a fim de comparar a eficácia das medidas de dissimilaridade propostas e também, no âmbito hierárquico, os métodos de ligação.

Com as aplicações, vimos que as técnicas de escalonamento multidimensional métrico e não métrico são de grande ajuda na visualização de grupos em dimensões menores, principalmente em $\mathbb{R}$ e $\mathbb{R}^2$, e também fornecem uma indicação da quantidade de grupos presentes no conjunto de dados. Além disso, quando trabalhando com dados métricos, a adição de um algoritmo de particionamento ao cMDS é uma excelente maneira de visualizar o agrupamento dos dados.

É importante ressaltar que, apesar de neste trabalho termos obtidos métodos de agrupamento, medidas de dissimilaridade e medidas de ligação mais eficazes nas aplicações realizadas, não podemos converter em regra, ou seja, não existe um método ou medida que seja melhor que outro em todos os aspectos.

Como sugestões para trabalhos futuros podem ser considerados outros métodos de agrupamento particionado, como K-medoides ou métodos *fuzzy*. Além disso, podem ser consideradas outras medidas de distância – preferencialmente para dados híbridos – ou outros métodos de ligação com a finalidade de expandir as possibilidades de agrupamento dos dados.

REFERÊNCIAS

BISHOP, Christopher M. **Pattern Recognition and Machine Learning**. London: Springer, 2006.

BORG, I; GROENEN, P. J. F. **Modern Multidimensional Scaling**. New York, NY: Springer New York, 2005. .978-0-387-25150-9.

DUDA, Richard O.; STORK, David G.; HART, Peter E. **Pattern Classification**. 2ª ed. [S.l.]: Wiley-Interscience, 2000.

FISHER, R. A. **The use of multiple measurements in taxonomic problems**. Annals of Eugenics v. 7, n. 2, set. 1936.

FOSS, Alexander H.; MARKATOU, Marianthi; RAY, Bonnie. **Distance Metrics and Clustering Methods for Mixed-type Data.** International Statistical Review v. 87, n. 1, 21 abr. 2019.

GERSHO, Allen; GRAY, Robert M. **Vector Quantization and Signal Compression**. Boston, MA: Springer US, 1992. .978-1-4613-6612-6.

GOWER, J. C. A **General Coefficient of Similarity and Some of Its Properties**. Biometrics v. 27, n. 4, dez. 1971.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2ª ed. [S.l.]: Springer-Verlag, 2009.

JAMES, Gareth *et al.* **An Introduction to Statistical Learning**. New York, NY: Springer New York, 2013. 103 v. .978-1-4614-7137-0.

KASSAMBARA, Alboukadel. **Practical Guide to Cluster Analysis in R**. 1. ed. [S.l.]: STHDA, 2017. Disponível em: http://www.sthda.com.

KAUFMAN, Leonard; ROUSSEEUW, Peter J. **Finding Groups in Data: An Introduction to Cluster Analysis**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 1990. .9780470316801.

KRUSKAL, J. B. **Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis**. Psychometrika v. 29, n. 1, mar. 1964.

MACQUEEN, J B. **Some Methods for Classification and Analysis of Multivariate Observations**. 1967, [S.l.]: University of California Press, 1967. p.281–297.

MARSLAND, Stephen. **Machine Learning: An Algorithmic Perspective**. 2ª ed. London: Chapman & Hall/CRC, 2014.

R CORE TEAM. *R:* **A Language and Environment for Statistical Computing**. Vienna, Áustria: [s.n.]. Disponível em: https://www.R-project.org/. 2020

SARLE, W S. **Cubic Clustering Criterion**. [S.l.]: SAS Institute, 1983. Disponível em: https://books.google.com.br/books?id=YynlGAAACAAJ. (SAS technical report).

TIBSHIRANI, Robert; WALTHER, Guenther; HASTIE, Trevor. **Estimating the number of clusters in a data set via the gap statistic**. Journal of the Royal Statistical Society: Series B (Statistical Methodology) v. 63, n. 2, 2001.

WARD, Joe H. **Hierarchical Grouping to Optimize an Objective Function**. Journal of the American Statistical Association v. 58, n. 301, mar. 1963.

APÊNDICE

Os comandos abaixo foram utilizados na realização deste trabalho. Este apêndice foi gerado em R Markdown e os pacotes utilizados foram: *knitr*, *tidyverse, cluster, factoextra, dendextend, useful, ggplot2, ggpubr, NbClust, MASS e stats*.

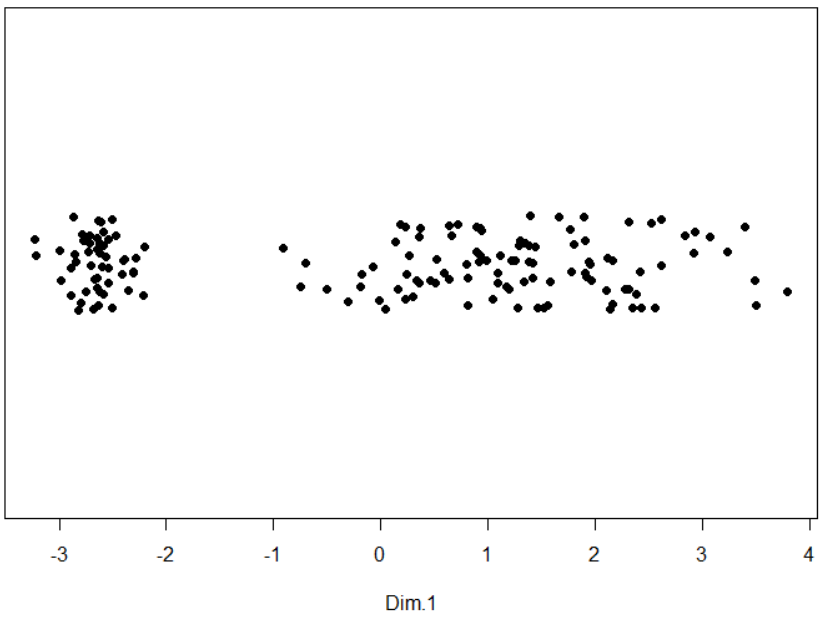
IRIS

O banco de dados IRIS fornece as medidas, em centímetros, das variáveis comprimento e largura tanto das pétalas quanto das sépalas, respectivamente, para 50 flores de cada uma das 3 espécies de íris.

Escalonamento Multidimensional Clássico (cMDS)

Uma dimensão

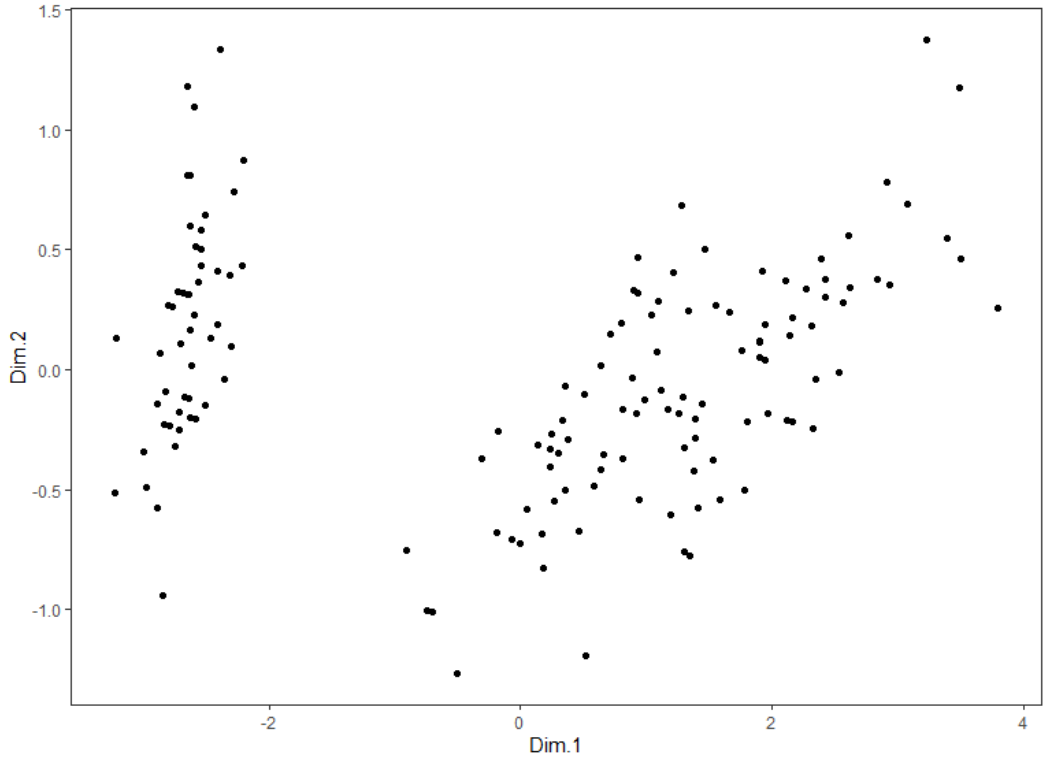
```
# Matriz de dissimilaridade
iris.df <- iris[,-5]
dist.obj <- get_dist(iris.df, method = "euclidean")
dist.matrix <- as.matrix(dist.obj)

# cMDS uma dimensão
mds <- cmdscale(dist.obj, eig=TRUE, k=1)

x <- mds$points[,1]
stripchart(x, pch = 19, xlab = "Dim.1", method = "jitter")
```




```
# Coeficiente de mensuração
sum(mds$eig[1])/sum(mds$eig)
## [1] 0.9246187
```

Duas dimensões

```
# cMDS duas dimensões
mds <- cmdscale(dist.obj,eig=TRUE, k=2)

x <- mds$points[,1]
y <- mds$points[,2]

ggplot(iris, aes(x=x, y=y)) +
  geom_point() +
  labs(x = "Dim.1", y = "Dim.2", title = "") +
  theme_bw() +
  theme(
    plot.background = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
  )
```




```
# Coeficiente de mensuração
sum(mds$eig[1:2])/sum(mds$eig)
## [1] 0.9776852
```

K-médias

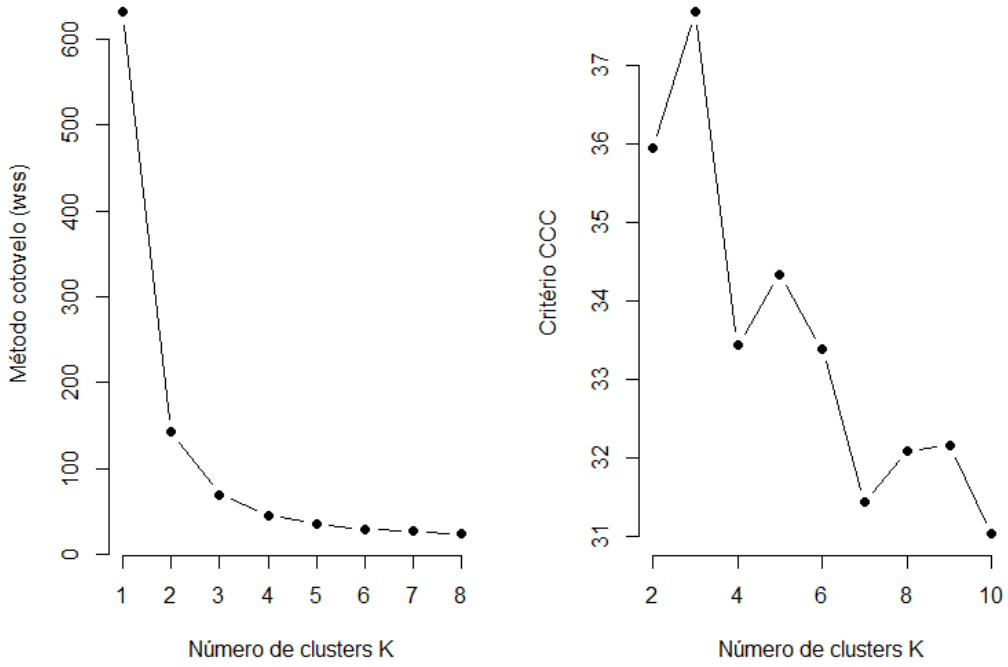
Escolha da medida de distância entre observações

```
# Escolha da medida de distância k = 2
iris.df <- scale(iris.df)

## EUCLIDIANA
data <- daisy(iris.df, metric = "euclidean")
k.eu <- kmeans(data, centers = 2, nstart = 25)
eucli2 <- k.eu$betweenss/k.eu$totss

## MANHATTAN
data <- daisy(iris.df, metric = "manhattan")
k.man <- kmeans(data, centers = 2, nstart = 25)
manh2 <- k.man$betweenss/k.man$totss

## GOWER
data <- daisy(iris.df, metric = "gower")
k.gower <- kmeans(data, centers = 2, nstart = 25)
gow2 <- k.gower$betweenss/k.gower$totss

# Escolha da medida de distância k = 3

## EUCLIDIANA
data <- daisy(iris.df, metric = "euclidean")
k.eu <- kmeans(data, centers = 3, nstart = 25)
eucli3 <- k.eu$betweenss/k.eu$totss

## MANHATTAN
data <- daisy(iris.df, metric = "manhattan")
k.man <- kmeans(data, centers = 3, nstart = 25)
manh3 <- k.man$betweenss/k.man$totss

## GOWER
data <- daisy(iris.df, metric = "gower")
k.gower <- kmeans(data, centers = 3, nstart = 25)
gow3 <- k.gower$betweenss/k.gower$totss
```

Dissimilaridade	R^2 (k = 2)	R^2 (k = 3)
Euclidiana	0.712	0.823
Manhattan	0.743	0.852
Gower	**0.775**	**0.889**

Escolha do número de grupos

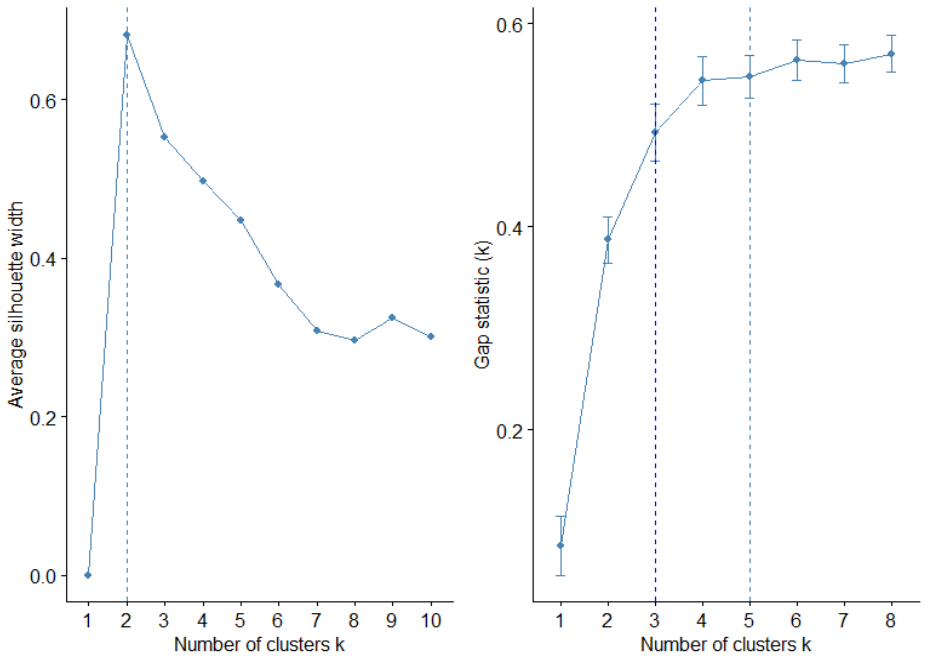
```
data <- daisy(iris.df, metric = "gower")

# Função para calcular o wss - método cotovelo
wss <- function(k) {
  kmeans(data, k, nstart = 25)$tot.withinss
}

par (mfrow = c(1,2))

# Método cotovelo
k.values <- 1:8
wss_values <- map_dbl(k.values, wss)
plot(k.values, wss_values,
     type="b", pch = 19, frame = FALSE,
     xlab="Número de clusters K",
     ylab="Método cotovelo (wss)")

# CCC
ccc <- NbClust(iris.df, min.nc=2, max.nc=10, index = "ccc", method = "
kmeans")
k.values <- 2:10
plot(k.values, as.numeric(ccc$All.index),
     type="b", pch = 19, frame = FALSE,
     xlab="Número de clusters K",
     ylab="Critério CCC")
```



```
# Silhueta média
silhueta <- fviz_nbclust(iris.df, kmeans, method = "silhouette") + ggt
itle ("")

# Estatística de lacuna
gap_stat <- clusGap(iris.df, FUN = kmeans, nstart = 25,
                    K.max = 8, B = 50)
print(gap_stat, method = "firstmax")
```

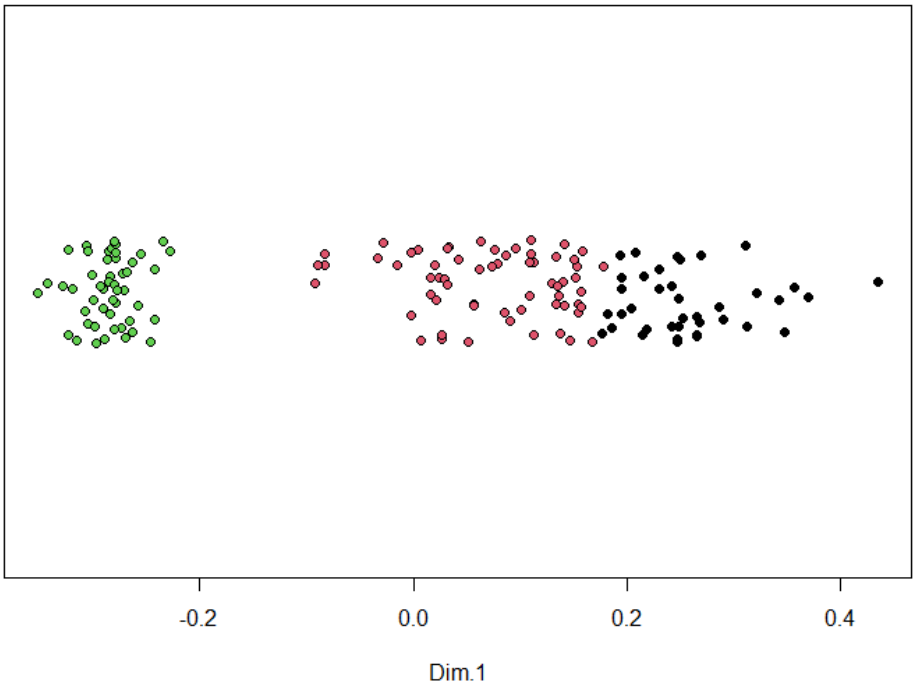
```
## Clustering Gap statistic ["clusGap"] from call:
## clusGap(x = iris.df, FUNcluster = kmeans, K.max = 8, B = 50, nstart
= 25)
## B=50 simulated reference sets, k = 1..8; spaceH0="scaledPCA"
##  --> Number of clusters (method 'firstmax'): 3
##          logW   E.logW       gap     SE.sim
## [1,] 4.534565 4.757985 0.2234204 0.02851807
## [2,] 4.021316 4.492107 0.4707910 0.02628340
## [3,] 3.806577 4.298117 0.4915395 0.02435556
## [4,] 3.699263 4.143565 0.4443016 0.02189701
## [5,] 3.589284 4.052873 0.4635892 0.02031706
## [6,] 3.522810 3.975813 0.4530032 0.02210836
## [7,] 3.448288 3.911136 0.4628478 0.02073384
## [8,] 3.379870 3.853027 0.4731572 0.01944001
```

```
lacuna <- fviz_gap_stat(gap_stat) +
  geom_vline(xintercept = 3, linetype = 2, col = "blue") +
  ggtitle ("")

gridExtra::grid.arrange(silhueta, lacuna, nrow = 1)
```



Complementando o cMDS de uma dimensão com o K-médias

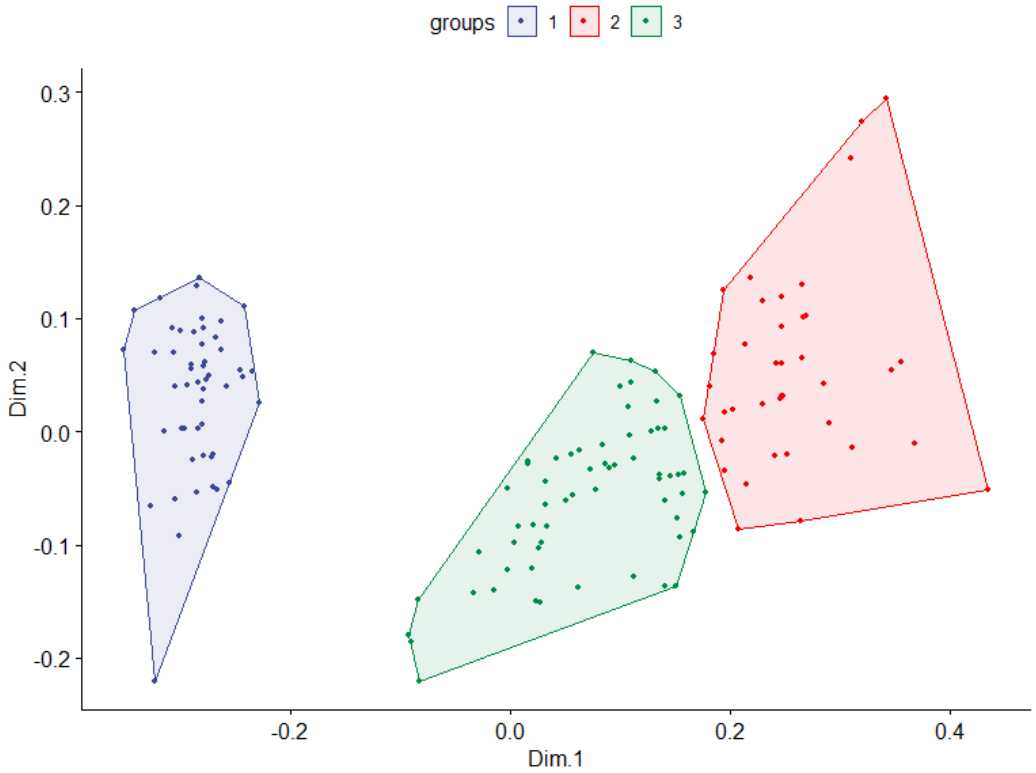
```
dist.obj <- get_dist(iris.df, method = "euclidean")
dist.matrix <- as.matrix(dist.obj)
mds <- cmdscale(data,eig=TRUE, k=1)

mds.df <- as.data.frame(mds$points)
colnames(mds.df) <- c("Dim.1")

data <- daisy(iris.df, metric = "gower")
k.gower <- kmeans(data, centers = 3, nstart = 25)
kmclusters <- as.factor(k.gower$cluster)
mds.df$groups <- kmclusters

with (mds.df,
      stripchart(Dim.1, pch = 21, xlab = "Dim.1", bg = groups, method
= "jitter"))
```



Complementando o cMDS de duas dimensões com o K-médias

```
dist.obj <- get_dist(iris.df, method = "euclidean")
dist.matrix <- as.matrix(dist.obj)
mds <- cmdscale(data,eig=TRUE, k=2)

mds.df <- as.data.frame(mds$points)
colnames(mds.df) <- c("Dim.1", "Dim.2")

data <- daisy(iris.df, metric = "gower")
k.gower <- kmeans(data, centers = 3, nstart = 25)
kmclusters <- as.factor(k.gower$cluster)
```

```
mds.df$groups <- kmclusters
ggscatter(mds.df,
          x = "Dim.1",
          y = "Dim.2",
          color = "groups",
          palette = "aaas",
          size = 1,
          ellipse = TRUE,
          ellipse.type = "convex",
          repel = TRUE,
          max.overlaps = 50)
```



Agrupamento hierárquico aglomerativo - AGNES

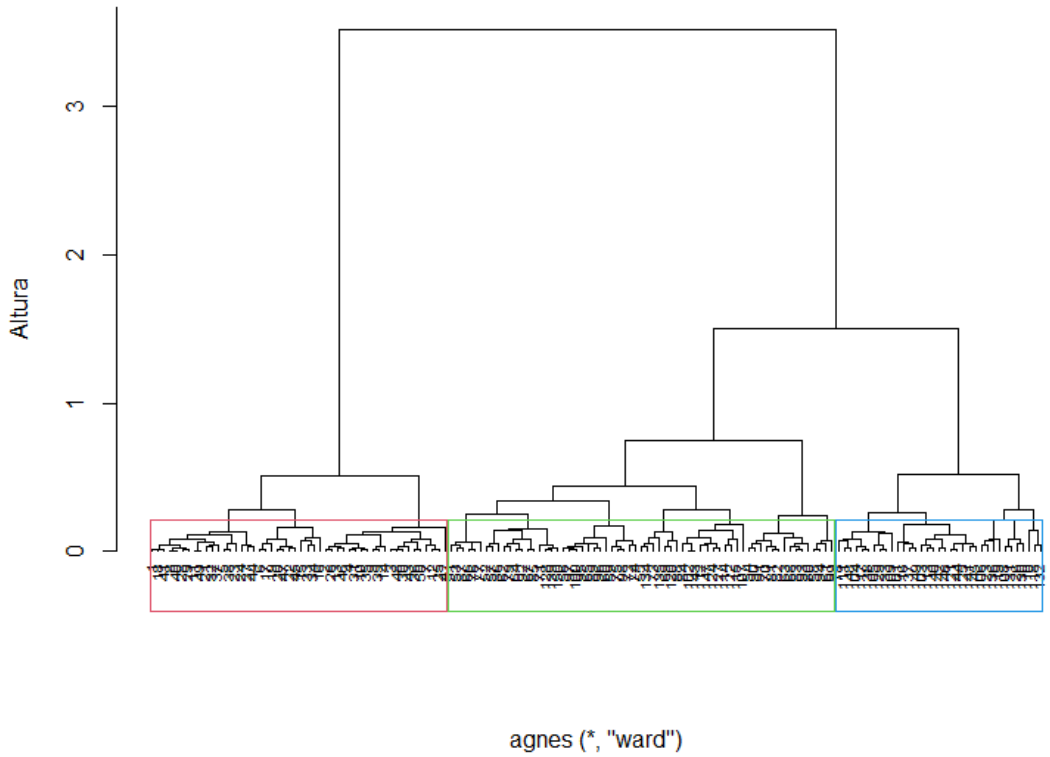
```
# métodos de ligação propostos
m <- c( "average", "single", "complete", "ward")
names(m) <- c( "average", "single", "complete", "ward")

# função para computar o coeficiente aglomerativo
d <- daisy(iris.df, metric = "gower")

ac <- function(x) {
  agnes(d, method = x)$ac
}
map_dbl(m, ac)
```

```
##  average    single    complete      ward
## 0.9267739  0.8549057  0.9576475  0.9901241
```

```
iris.agnes <- agnes(d, method = "ward")

# tamanho dos grupos
sub_ward <- cutree(iris.agnes, k = 3)
table(sub_ward)
```

```
## sub_ward
##  1  2  3
## 50 65 35
```

```
# dendrograma AGNES; k = 3
pltree(iris.agnes, cex = 0.6, hang = -1, main = "", xlab = "", ylab =
"Altura")
rect.hclust(iris.agnes, k = 3, border = 2:5)
```



Agrupamento hierárquico divisivo - DIANA

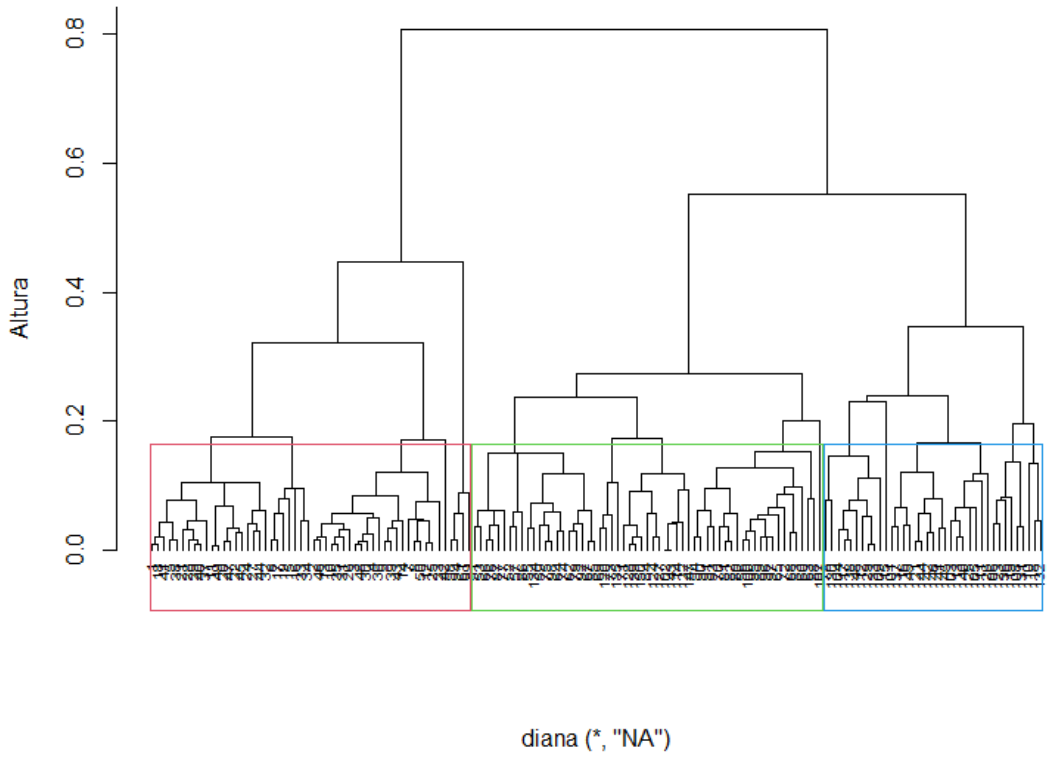
```
# Não precisamos escolher método de ligação
iris.diana <- diana(d)

# Coeficiente divisivo
iris.diana$dc # menor que o coeficiente aglomerativo anterior -> menos
eficácia
```

```
## [1] 0.9502729
```

```
sub_diana <- cutree(iris.diana, k = 3)
table(sub_diana)
```

```
## sub_diana
##  1  2  3
## 54 59 37
```

```
# Dendrograma DIANA; k = 3
pltree(iris.diana, cex = 0.6, hang = -1, main = "", xlab = "", ylab =
"Altura")
rect.hclust(iris.diana, k = 3, border = 2:5)
```



SEGMENTAÇÃO DE CLIENTES

O banco de dados do shopping fornece as características de 200 clientes membros, tais como: sexo, idade, renda anual (em milhares) e pontuação de gasto.

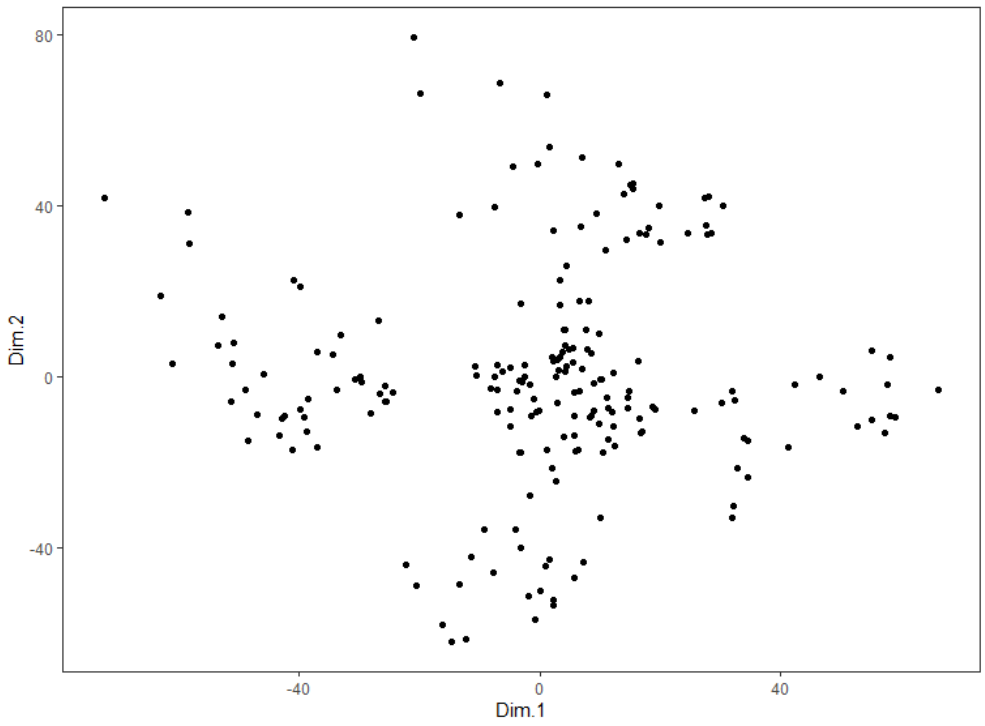
Escalonamento Multidimensional não métrico (nMDS)

Duas dimensões

```
# nMDS duas dimensões
distancias <- daisy(data, metric = "euclidean")
mds <- isoMDS(distancias, k = 2)
mds$stress #razoável
```

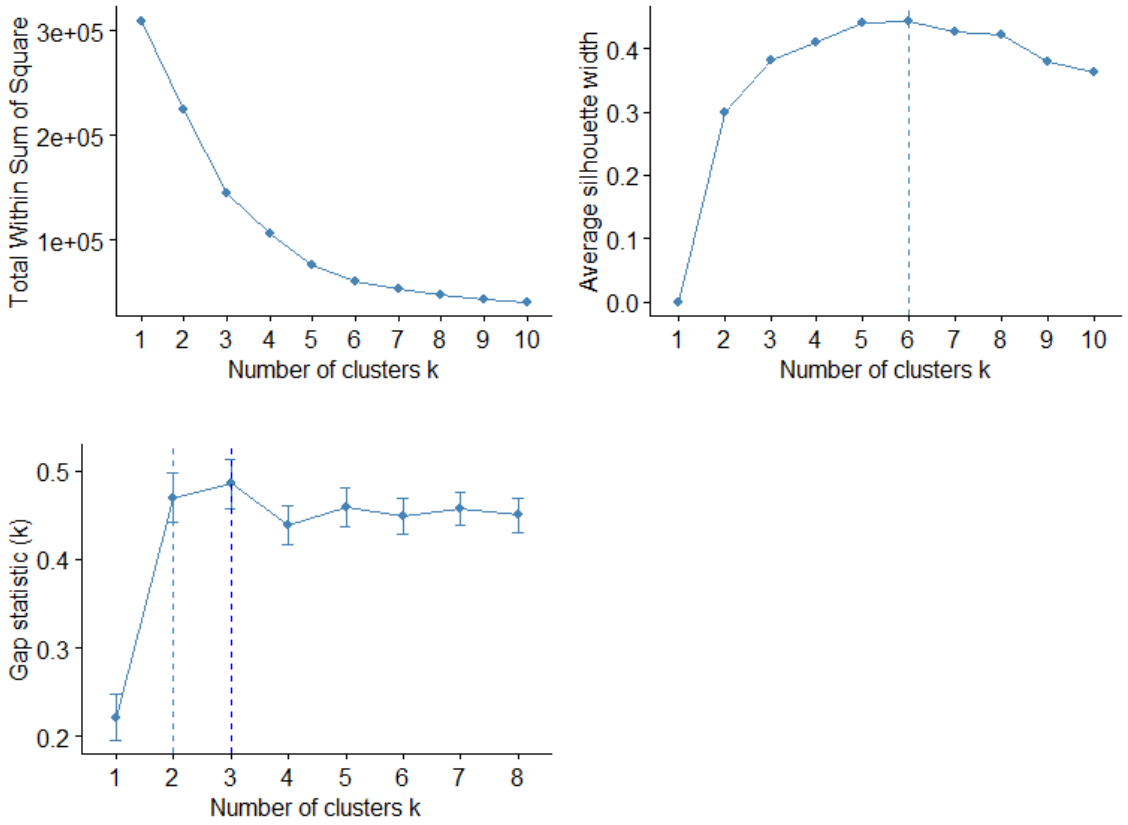
```
## [1] 11.83113
```

```
x <- mds$points[,1]
y <- mds$points[,2]

ggplot(data, aes(x=x, y=y)) +
  geom_point() +
  labs(x = "Dim.1", y = "Dim.2", title = "") +
  theme_bw() +
  theme(
    plot.background = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
  )
```

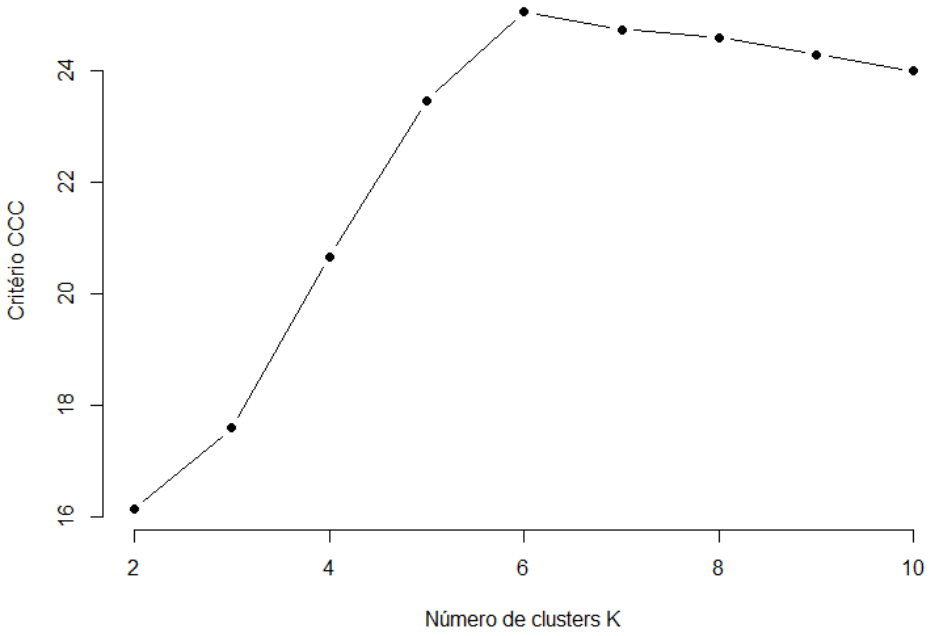


Escolha do número de grupos

```
# Método cotovelo
p1<- fviz_nbclust(data, FUN = hcut, method = "wss") + ggtitle("")

# Método da silhueta média
p2 <- fviz_nbclust(data, FUN = hcut, method = "silhouette") + ggtitle
("")

# Método da estatística de lacuna
gap_stat <- clusGap(data, FUN = hcut, nstart = 25, K.max = 10, B = 50)
p3 <- fviz_gap_stat(gap_stat) +
  geom_vline(xintercept = 6, linetype = 2, col = "blue") +
  ggtitle ("")

gridExtra::grid.arrange(p1, p2, p3,nrow = 2)
```



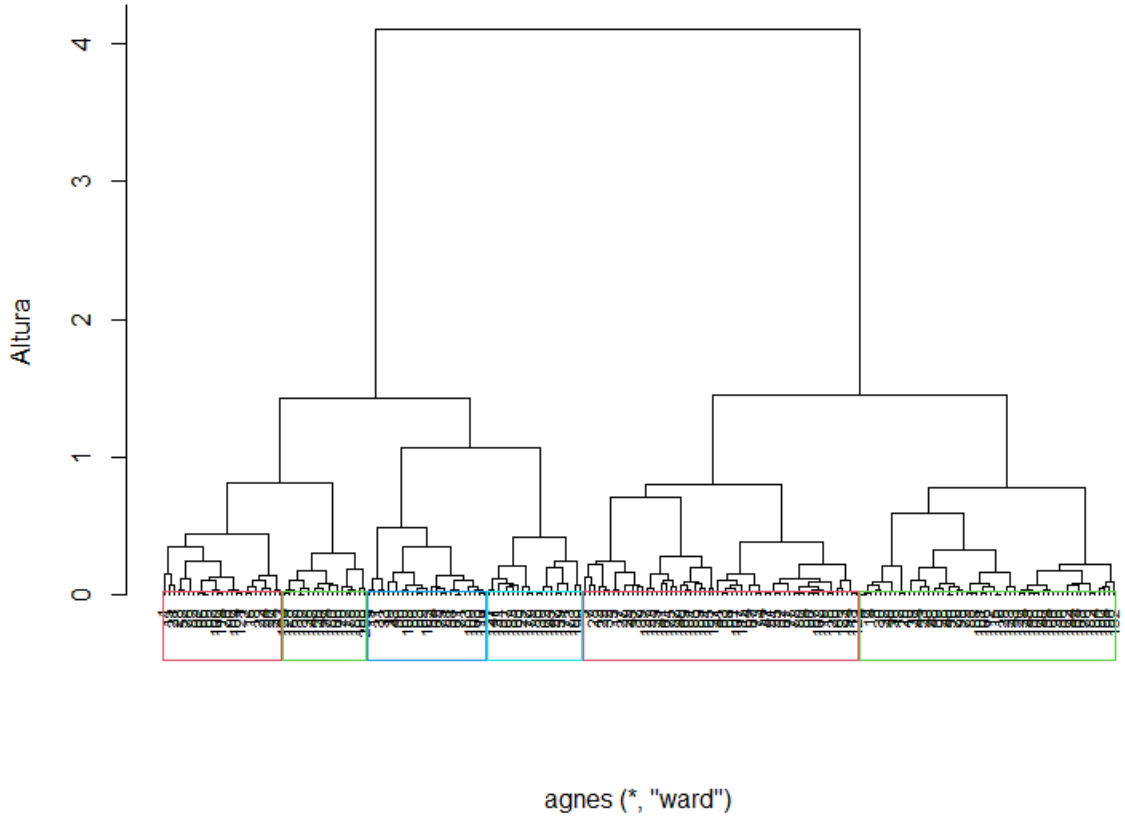

```
# CCC
ccc <- NbClust(data, min.nc=2, max.nc=10, index = "ccc", method = "war
d.D2")
k.values <- 2:10
plot(k.values, as.numeric(ccc$All.index),
     type="b", pch = 19, frame = FALSE,
     xlab="Número de clusters K",
     ylab="Critério CCC")
```



Agrupamento hierárquico aglomerativo - AGNES

```
# métodos de ligação propostos
m <- c( "average", "single", "complete", "ward")
names(m) <- c( "average", "single", "complete", "ward")

# função para computar o coeficiente aglomerativo
d <- daisy(data, metric = "gower")
## Warning in daisy(data, metric = "gower"): binary variable(s) 1 trea
ted as
## interval scaled
ac <- function(x) {
  agnes(d, method = x)$ac
}

map_dbl(m, ac)
##   average   single  complete      ward
## 0.9255152 0.8802714 0.9559780 0.9914272

market.agnes <- agnes(d, method = "ward")

# tamanho dos grupos
sub_ward <- cutree(market.agnes, k = 6)
table(sub_ward)
## sub_ward
##  1  2  3  4  5  6
## 25 58 54 25 18 20
```

```
# dendrograma AGNES; k = 6
pltree(market.agnes, cex = 0.6, hang = -1, main = "", xlab = "", ylab
= "Altura")
rect.hclust(market.agnes, k = 6, border = 2:5)
```



Agrupamento hierárquico divisivo - DIANA

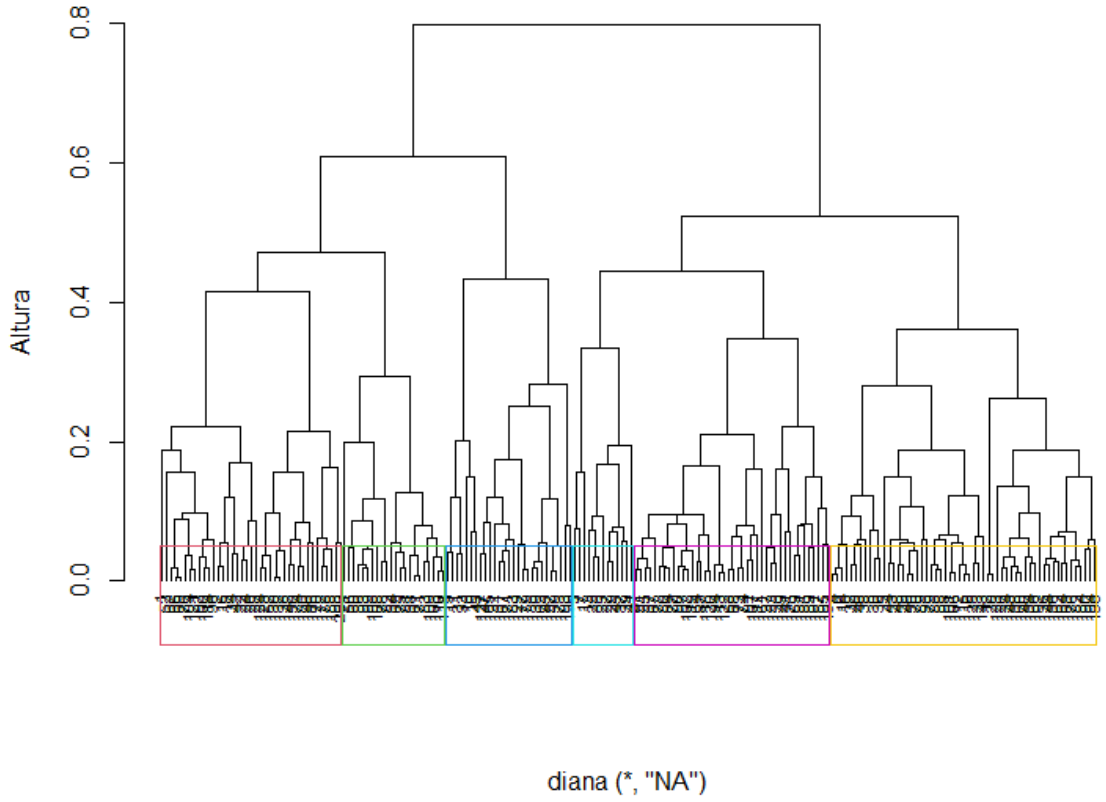
```
# Não precisamos escolher método de ligação
market.diana <- diana(d)

# Coeficiente divisivo
market.diana$dc
```

```
## [1] 0.9485657
```

```
# tamanho dos grupos
sub_diana <- cutree(market.diana, k = 6)
table(sub_diana)
```

```
## sub_diana
##  1  2  3  4  5  6
## 39 13 57 27 22 42
```

```
# Dendrograma DIANA; k = 6
pltree(market.diana, cex = 0.6, hang = -1, main = "", xlab = "", ylab
= "Altura")
rect.hclust(market.diana, k = 6, border = 2:8)
```



Resumo das variáveis nos grupos criados

AGNES

```
data %>%
  mutate(Cluster = sub_ward) %>%
  group_by(Cluster) %>%
  summarise_all("mean")
```

```
## # A tibble: 6 x 5
##   Cluster  Sexo Idade `Renda(k)` Score_Gasto
##     <int> <dbl> <dbl>      <dbl>       <dbl>
## 1       1     0  25.7       40.4        59
## 2       2     1  47.4       58.3        35.7
## 3       3     1  28.1       60.3        68.5
## 4       4     0  58.8       47.8        41
## 5       5     0  33.3       87.1        82.7
## 6       6     0  39.5       85.2        14.0
```

DIANA

```
data %>%
  mutate(Cluster = sub_diana) %>%
  group_by(Cluster) %>%
  summarise_all("mean")
```

```
## # A tibble: 6 x 5
##   Cluster  Sexo Idade `Renda(k)` Score_Gasto
##     <int> <dbl> <dbl>      <dbl>       <dbl>
## 1       1     0  28         62.2        72.1
## 2       2     1  44.2       28.2        20.3
## 3       3     1  28.4       59.7        67.7
## 4       4     0  43.4       69.9        14.2
## 5       5     0  56.4       52.9        48.8
## 6       6     1  49.3       68.3        39.3
```