

Big Data

Tema da aula
Análise de Cluster



BUSINESS SCHOOL

Graduação, pós-graduação, MBA, Pós-MBA, Mestrado Profissional, Curso In Company e EAD



CONSULTING

Consultoria personalizada que oferece soluções baseada em seu problema de negócio



RESEARCH

Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino



Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais **escolas de negócio do mundo**, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 **projetos de consultorias** em organizações públicas e privadas.



Único curso de graduação em administração a receber as notas máximas



A primeira escola brasileira a ser finalista da maior competição de MBA do mundo



Única *Business School* brasileira a figurar no *ranking* LATAM



Signatária do Pacto Global da ONU



Membro fundador da ANAMBA - Associação Nacional MBAs



Credenciada pela AMBA - Association of MBAs



Credenciada ao Executive MBA Council



Filiada a AACSB - Association to Advance Collegiate Schools of Business



Filiada a EFMD - European Foundation for Management Development



Referência em cursos de MBA nas principais mídias de circulação

O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de **Big Data**, **Analytics** e **Inteligência Artificial**.



O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil

Os diretores foram professores de grandes especialistas do mercado

+10 anos de atuação

+1000 alunos formados

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria
- Larga experiência de mercado na resolução de *cases*
- Participação em Congressos Nacionais e Internacionais
- Professor assistente que acompanha o aluno durante todo o curso

Estrutura

- 100% das aulas realizadas em laboratórios
- Computadores para uso individual durante as aulas
- 5 laboratórios de alta qualidade (investimento +R\$2MM)
- 2 Unidades próximas a estação de metrô (com estacionamento)

Conteúdo da Aula

- 1. Introdução
 - i. Distância Euclidiana
- 2. Método Hierárquico
 - i. *Single* (vizinho mais próximo)
 - ii. *Complete* (vizinho mais longe)
- 3. Padronização de variáveis
 - i. Z-score
- 4. Método de Partição: K-médias
- 5. Exercícios



1. Introdução



Case: Encarteiramento de clientes

1. INTRODUÇÃO | ANÁLISE DE CLUSTER

6

Exemplo

Criar encarteiramento de clientes de um banco para atendimento diferenciado de acordo com investimento e relacionamento com o banco.

Aplicação

Segmento Bancário



Case: Canais de Atendimento

1. INTRODUÇÃO | ANÁLISE DE CLUSTER

7

Exemplo

Atendimento diferenciado no *call center* e centrais de atendimento.

Aplicação

SAC e Ouvidoria



Case: *People Analytics* - RH

1. INTRODUÇÃO | ANÁLISE DE CLUSTER

8

Exemplo

Estratégia de benefícios diferenciados de acordo com o estágio de vida dos funcionários de uma empresa.

Aplicação

Gestão de Pessoas



Case: Hábitos Alimentares

1. INTRODUÇÃO | ANÁLISE DE CLUSTER

9

Exemplo

Agrupar regiões com hábitos alimentares semelhantes e fazer um estudo em relação a longevidade e indicadores de saúde.

Aplicação

Áreas de Saúde & Nutrição



Case: CRM

1. INTRODUÇÃO | ANÁLISE DE CLUSTER

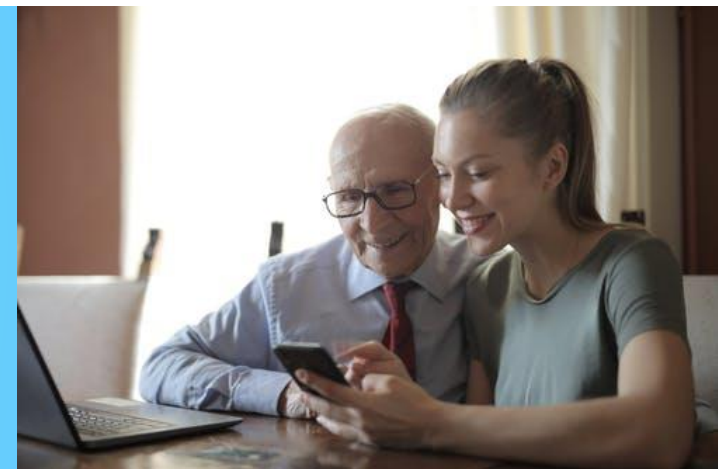
10

Exemplo

Segmentar clientes de acordo com o seu perfil sociodemográfico para comunicação de marketing de relacionamento diferenciado.

Aplicação

Área de Marketing e Comunicação



Case: Reconhecimento de Clientes

1. INTRODUÇÃO | ANÁLISE DE CLUSTER

11

Exemplo

Estratégia de reconhecimento e relacionamento com clientes de acordo com sua transacionalidade.

Aplicação

Marketing & CRM



Case: Varejo RFV

1. INTRODUÇÃO | ANÁLISE DE CLUSTER

12

Exemplo

Estratégia de reconhecimento e relacionamento com clientes de acordo com sua transacionalidade, baseada em Recência, Frequência e Valor.

Aplicação

Marketing & CRM

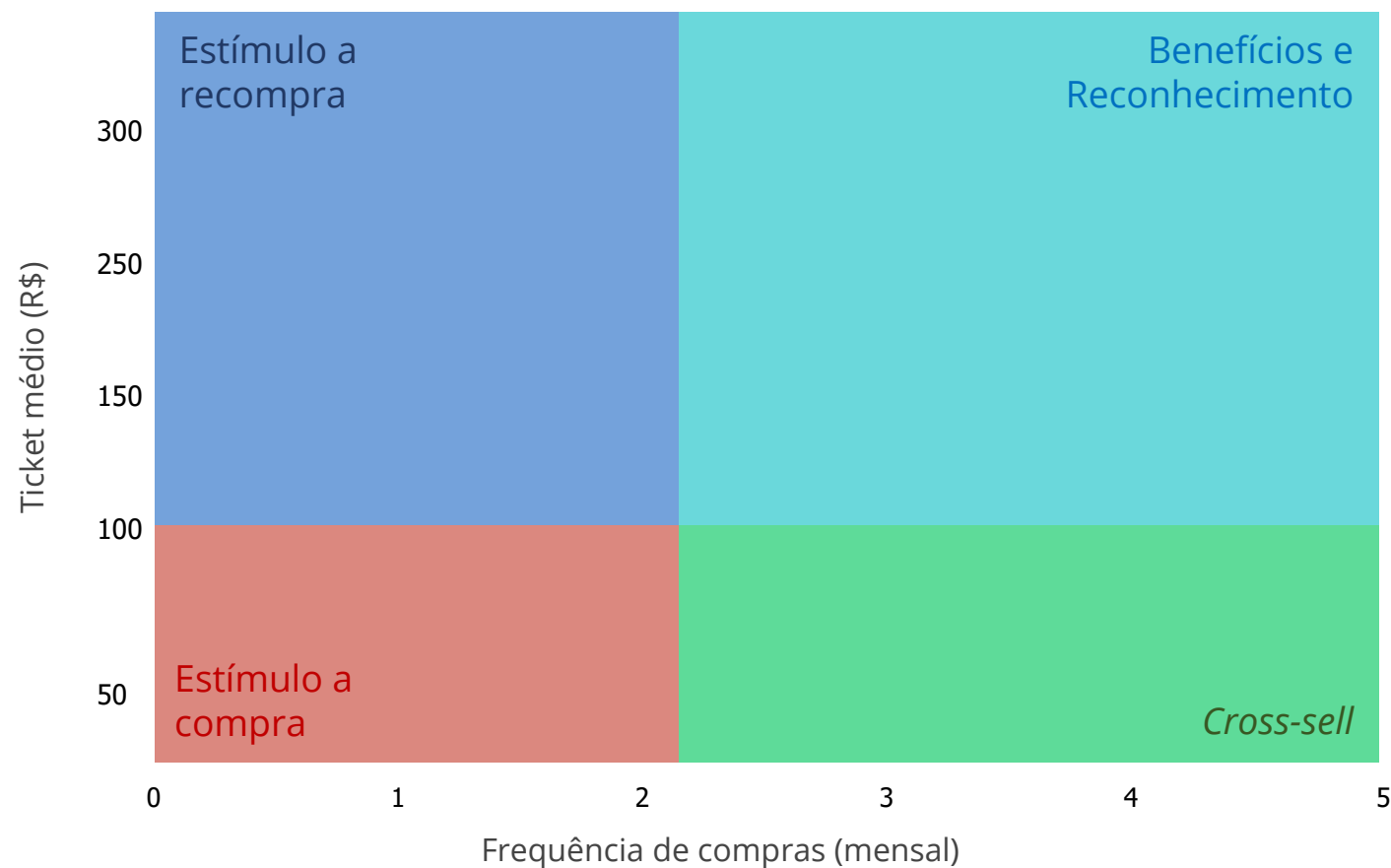


Case: Varejo (Frequência e Valor)

1. INTRODUÇÃO | ANÁLISE DE CLUSTER

13

Estratégias de reconhecimento e relacionamento segmentadas para **4 grupos** de transacionalidade, baseados em **frequência** e **valor**.

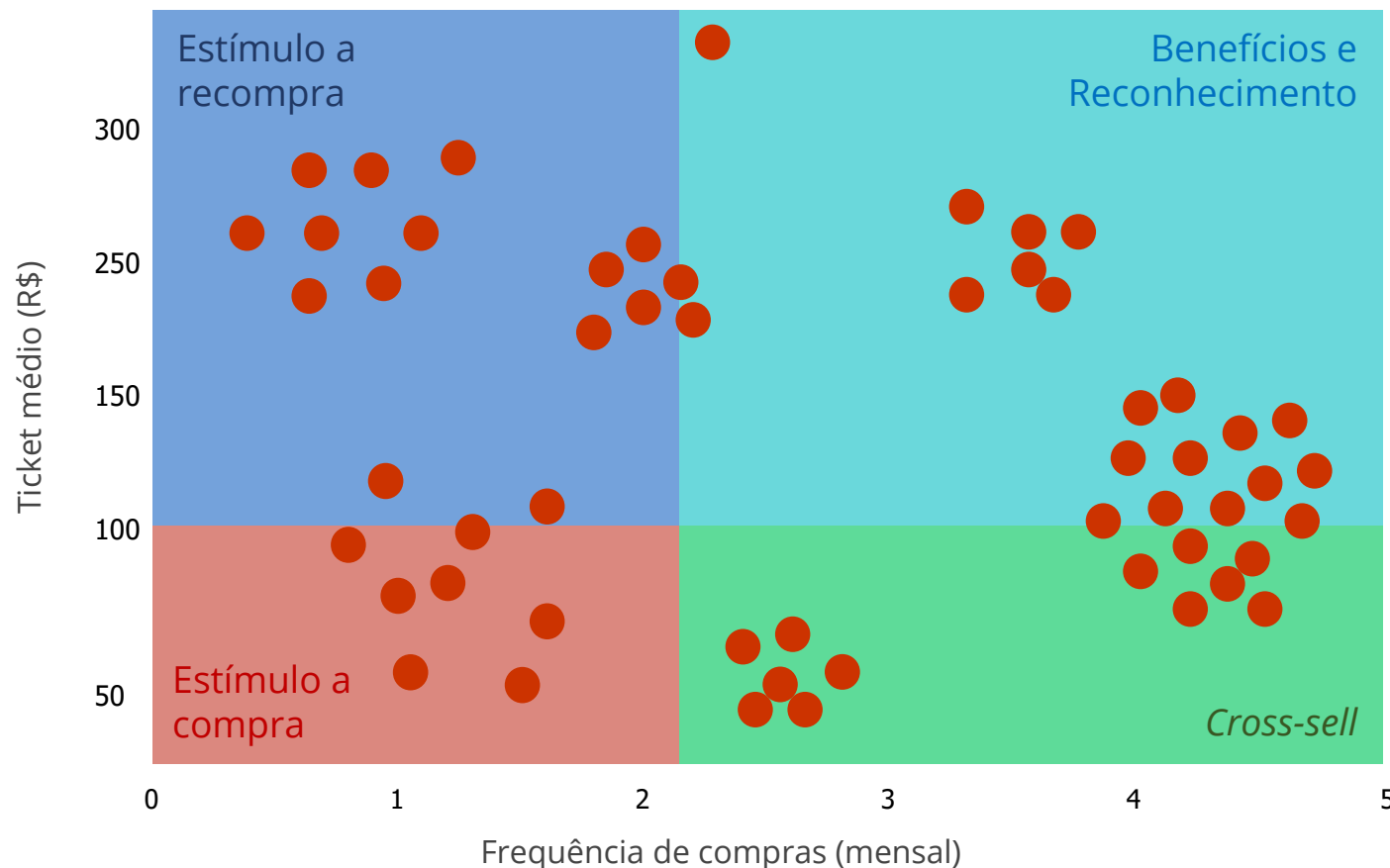


Case: Varejo (Frequência e Valor)

1. INTRODUÇÃO | ANÁLISE DE CLUSTER

14

Estratégias de reconhecimento e relacionamento segmentadas para **4 grupos** de transacionalidade, baseados em **frequência** e **valor**.



Uma segmentação baseada em **critérios de negócios** nem sempre fornece a melhor "regra" que agrupe os indivíduos semelhantes.

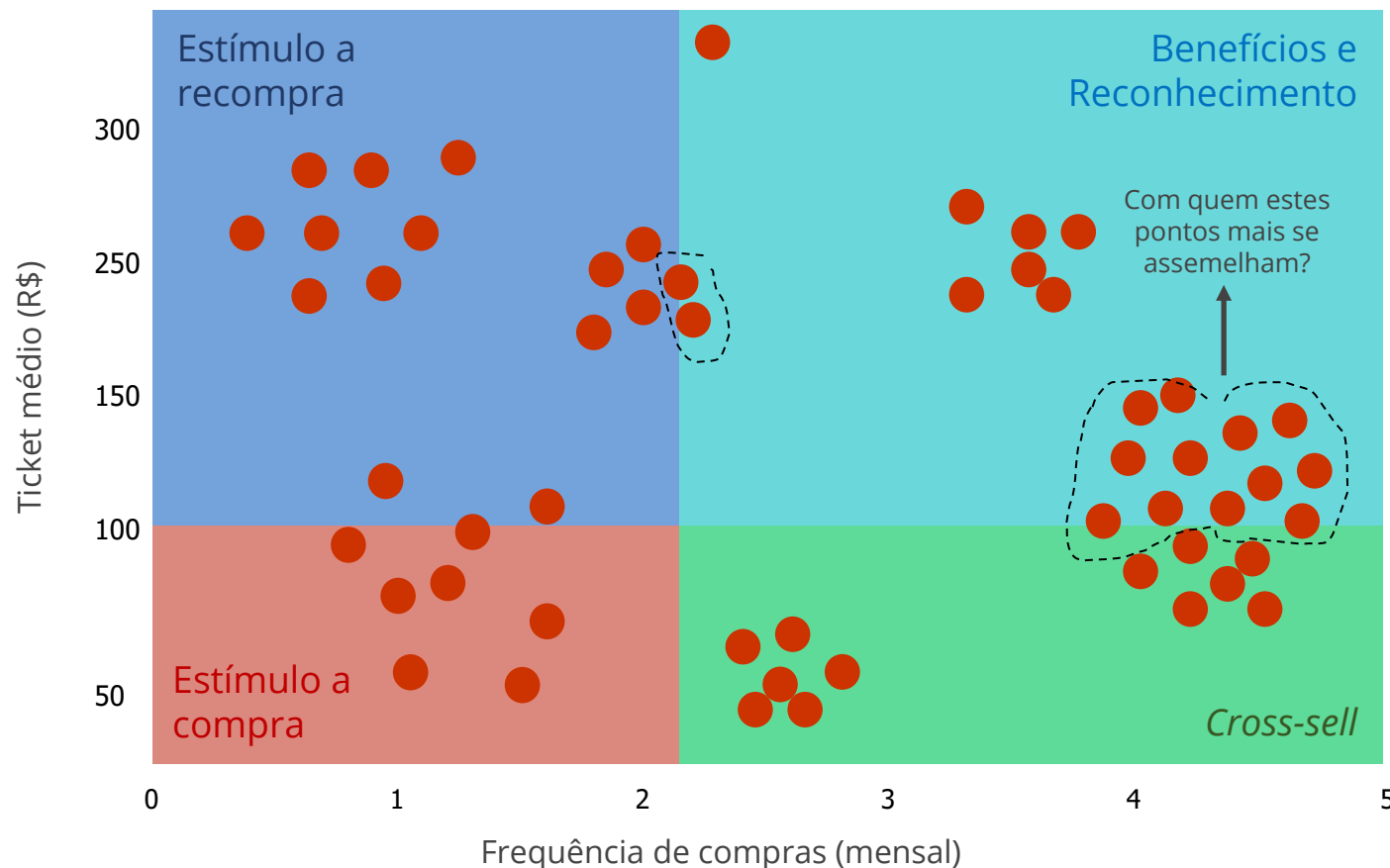


Case: Varejo (Frequência e Valor)

1. INTRODUÇÃO | ANÁLISE DE CLUSTER

15

Estratégias de reconhecimento e relacionamento segmentadas para **4 grupos** de transacionalidade, baseados em **frequência** e **valor**.

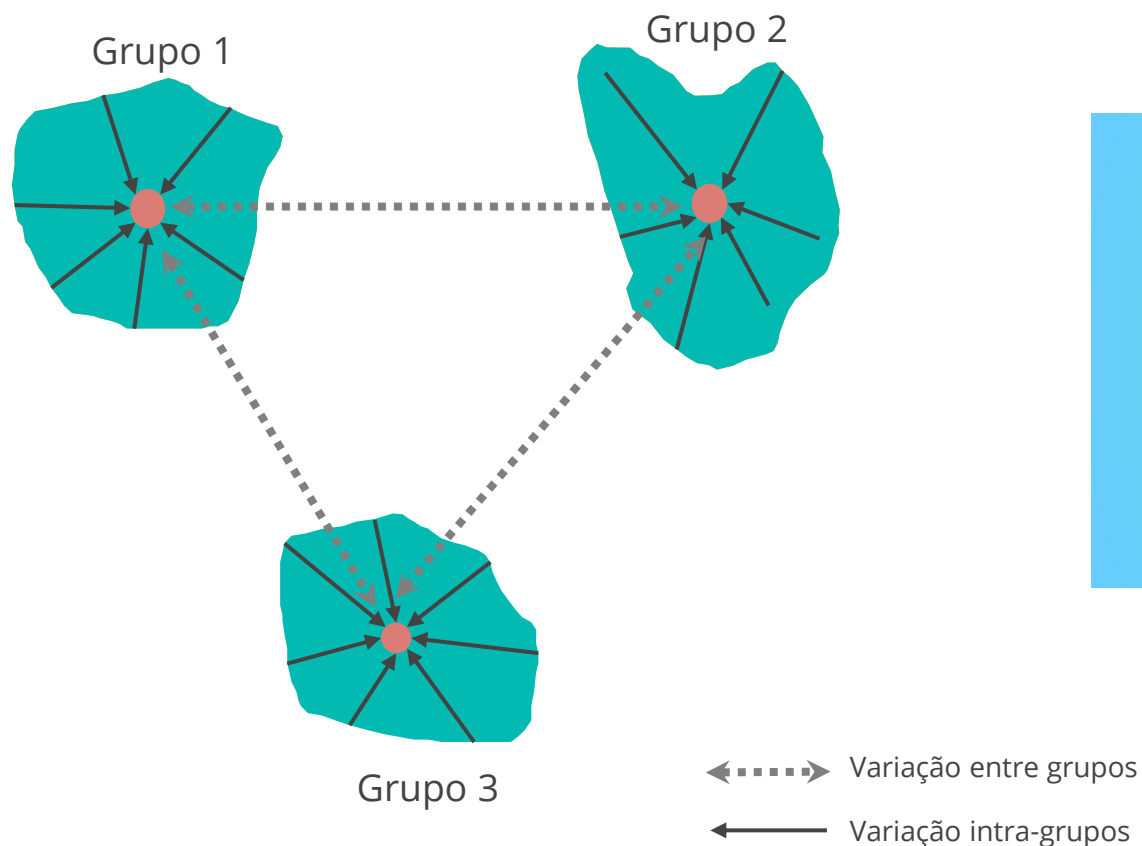


Uma segmentação baseada em **critérios de negócios** nem sempre fornece a melhor "regra" que agrupe os indivíduos semelhantes.

Objetivo da Análise de *Cluster*

1. INTRODUÇÃO | ANÁLISE DE CLUSTER

16



O objetivo da análise de *cluster* é **agrupar as observações** de tal forma que dentro de cada grupo as observações sejam **homogêneas entre si**; e os grupos sejam **heterogêneos**.

Desta forma, **dentro** de cada grupo a variabilidade deve ser **mínima**; e **entre** os grupos a variabilidade deve ser **máxima**.





Todos os exemplos citados anteriormente trazem aplicações práticas do uso da técnica de *Análise de Cluster* para **segmentar** públicos diferentes.

- Como definir as variáveis?
- Será que o modelo seleciona as características mais importantes?



Como identificamos indivíduos (observações) semelhantes?

1. INTRODUÇÃO | ANÁLISE DE CLUSTER

18

O **tigre** é mais parecido com o **gato** ou o **leão**?



Como identificamos indivíduos (observações) semelhantes?

1. INTRODUÇÃO | ANÁLISE DE CLUSTER

19

A semelhança entre os indivíduos dependerá da **variável de interesse**: porte ou elementos da face?



Porte



Elementos
da face





A parte mais difícil de um projeto que envolve Análise de *Cluster* é definir as variáveis, pois como é um método que não envolve variável resposta, **não há um critério de seleções de variáveis.**

Portanto, quem deve definir o objetivo é a área de negócios, e o especialista de análise de dados deve ter a habilidade de transformar os objetivos (informações de negócio) em variáveis para o algoritmo.



Segmentação

1. INTRODUÇÃO | ANÁLISE DE CLUSTER

21

Uma vez definidas quais as características que gostaríamos de avaliar como 'semelhantes', é necessária uma medida para **quantificar** essa semelhança.



Distância Euclidiana

1.i. DISTÂNCIA EUCLIDIANA | ANÁLISE DE CLUSTER

22

Considere o exemplo de uma analista de gestão de pessoas que deseja segmentar os candidatos em três grupos, considerando duas variáveis:

- a) tempo de formação do candidato (em anos);
- b) tempo que o candidato permaneceu na empresa anterior (em anos).

A tabela abaixo apresenta os valores das variáveis para 5 candidatos.

Candidato	Tempo de formação	Tempo na empresa anterior
1	2	2
2	3	4
3	12	12
4	8	16
5	12	2



Distância Euclidiana

1.i. DISTÂNCIA EUCLIDIANA | ANÁLISE DE CLUSTER

23

Na análise de *Cluster*, as observações são agrupadas de acordo com **medidas de dissimilaridade**.

Um critério de dissimilaridade que pode ser considerado para agrupar observações é a **Distância Euclidiana**. Quanto **menor** o seu valor, mais **parecidos** os elementos comparados.

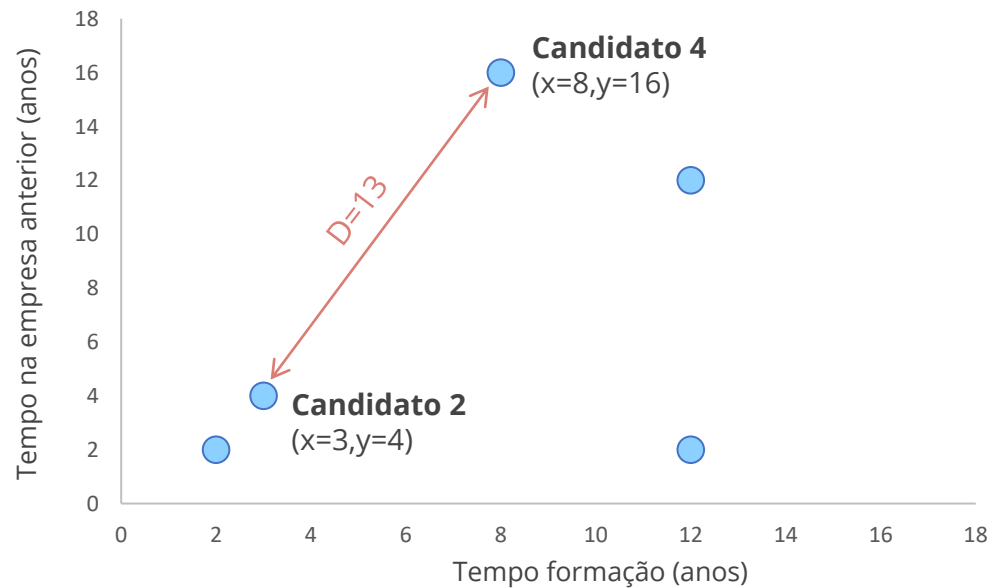


Exercício: Calcule a Distância Euclidiana

1.1. DISTÂNCIA EUCLIDIANA | ANÁLISE DE CLUSTER

24

Quem está mais próximo do candidato 2? O candidato 1 ou 4?



A **Distância Euclidiana (D)** entre os candidatos 2 e 4 é dada pela **reta vermelha**, e calculada por:

$$D^2 = (8 - 3)^2 + (16 - 4)^2 = 5^2 + 12^2 = 169$$

$$D = \sqrt{169} = 13$$

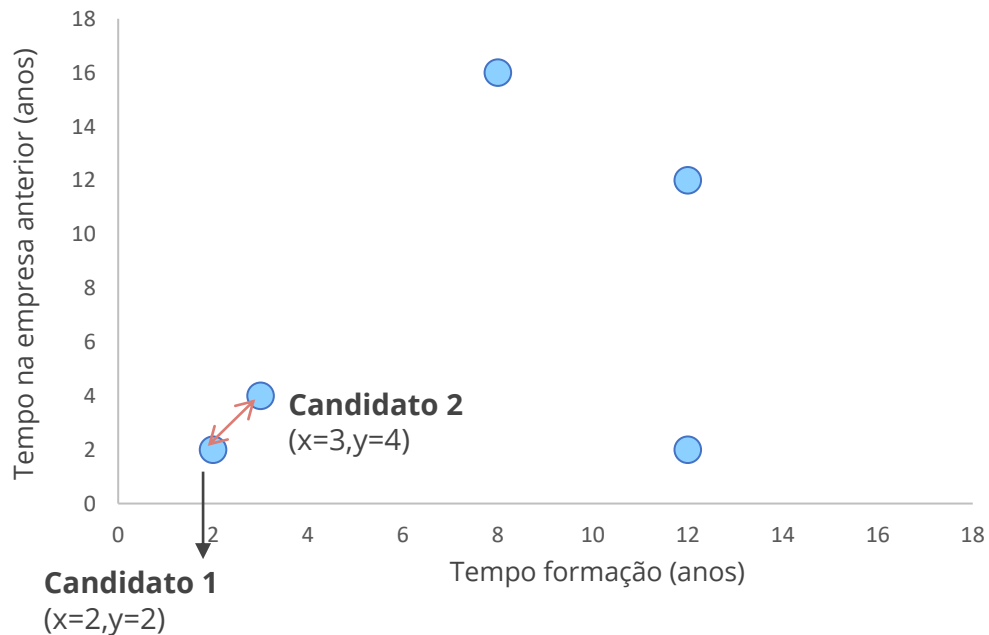


Exercício: Calcule a Distância Euclidiana

1.i. DISTÂNCIA EUCLIDIANA | ANÁLISE DE CLUSTER

25

Quem está mais próximo do candidato 2? O candidato 1 ou 4?



A **Distância Euclidiana (D)** entre os candidatos 1 e 2 é dada pela **reta vermelha**, e calculado por:

$$D^2 = (2 - 3)^2 + (2 - 4)^2 = (-1)^2 + (-2)^2 = 5$$

$$D = \sqrt{5} = 2,24$$

Quanto menor a distância, mais próximos os candidatos estão. Logo, o candidato 2 está mais próximo do candidato 1 e mais distante do candidato 4.



Matriz de distâncias

1.1. DISTÂNCIA EUCLIDIANA | ANÁLISE DE CLUSTER

26

Fazendo o cálculo das distâncias euclidianas entre todas as observações, obtém-se uma **matriz de distância**, que é **simétrica**.

Matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					



Matriz de distâncias

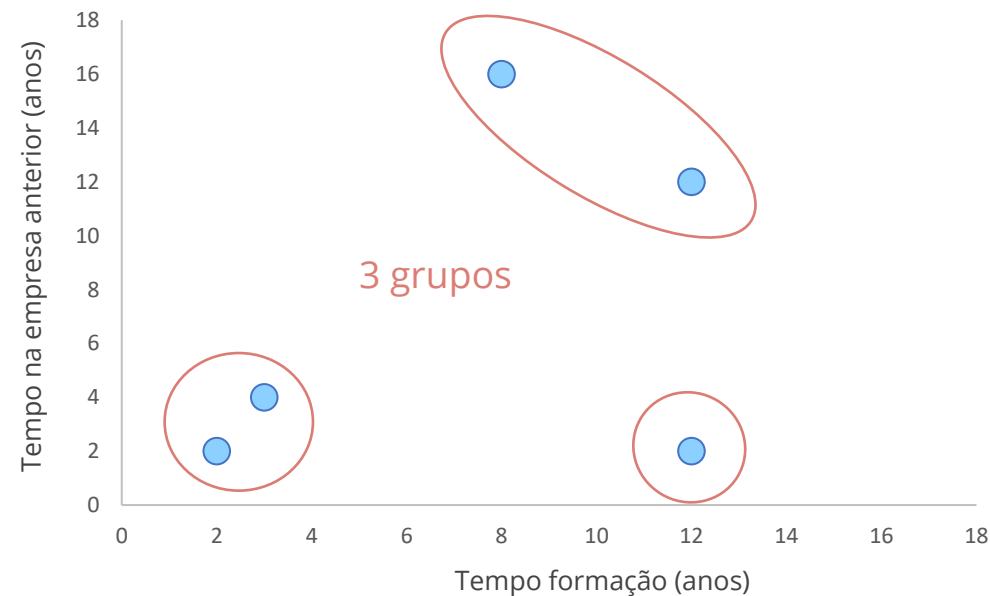
1.i. DISTÂNCIA EUCLIDIANA | ANÁLISE DE CLUSTER

27

Pela **matriz de distâncias**, pode-se observar quais elementos estão mais próximos (quanto menor a distância, mais próximos). Graficamente, é possível verificar a proximidade entre os candidatos.

Matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					



Como chegamos nestes 3 grupos?

2. Método Hierárquico



Discussão entre os métodos

2. MÉTODO HIERÁRQUICO | 2 MÉTODOS

29

- **Single** (vizinho mais próximo)
- **Complete** (vizinho mais longe)



Método *Single* (vizinho mais próximo)

2.i. MÉTODO SINGLE | ANÁLISE DE CLUSTER

30

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					



Método *Single* (vizinho mais próximo)

2.i. MÉTODO SINGLE | ANÁLISE DE CLUSTER

31

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1: juntar 1 e 2

Menor distância

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					



Método *Single* (vizinho mais próximo)

2.i. MÉTODO SINGLE | ANÁLISE DE CLUSTER

32

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1: juntar 1 e 2

Menor distância

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Agregamos pelo MÍNIMO

Dado que a menor distância é **2,24**, vamos agrupar as informações dos candidatos 1 e 2, por meio das **distâncias mínimas** em relação aos demais candidatos:

Distância entre 1 e 3 = 14,14

Distância entre 2 e 3 = 12,04

MÍNIMO é 12,04

Distância entre 1 e 4 = 15,23

Distância entre 2 e 4 = 13,00

MÍNIMO é 13,00

Distância entre 1 e 5 = 10,00

Distância entre 2 e 5 = 9,22

MÍNIMO é 9,22



Método *Single* (vizinho mais próximo)

2.i. MÉTODO SINGLE | ANÁLISE DE CLUSTER

33

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1: juntar 1 e 2

Menor distância

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Agregação pelo mínimo

	1 + 2	3	4	5
1 + 2		12,04	13,00	9,22
3			5,66	10,00
4				14,56
5				



Método *Single* (vizinho mais próximo)

2.i. MÉTODO SINGLE | ANÁLISE DE CLUSTER

34

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1: juntar 1 e 2

Menor distância

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 2: juntar 3 e 4

Menor distância e
agregação pelo mínimo

	1 + 2	3	4	5
1 + 2		12,04	13,00	9,22
3			5,66	10,00
4				14,56
5				



Método *Single* (vizinho mais próximo)

2.i. MÉTODO SINGLE | ANÁLISE DE CLUSTER

35

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1: juntar 1 e 2

Menor distância

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 2: juntar 3 e 4

Menor distância e
agregação pelo mínimo

	1 + 2	3	4	5
1 + 2		12,04	13,00	9,22
3			5,66	10,00
4				14,56
5				

Agregamos
pelo MÍNIMO

Dado que a menor distância é **5,66**, vamos agrupar as informações dos candidatos 3 e 4, por meio das **distâncias mínimas** em relação aos demais candidatos:

Distância entre 3 e 1+2 = 12,04

Distância entre 4 e 1+2 = 13,00

MÍNIMO é 12,04

Distância entre 3 e 5 = 10,00

Distância entre 4 e 5 = 14,56

MÍNIMO é 10,00



Método *Single* (vizinho mais próximo)

2.i. MÉTODO SINGLE | ANÁLISE DE CLUSTER

36

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

↓
Agregação pelo mínimo

	1 + 2	3 + 4	5
1 + 2		12,04	9,22
3 + 4			10,00
5			

Passo 1: juntar 1 e 2

↓
Menor distância

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 2: juntar 3 e 4

↓
Menor distância e
agregação pelo mínimo

	1 + 2	3	4	5
1 + 2		12,04	13,00	9,22
3			5,66	10,00
4				14,56
5				



Método *Single* (vizinho mais próximo)

2.i. MÉTODO SINGLE | ANÁLISE DE CLUSTER

37

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1: juntar 1 e 2

Menor distância

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 2: juntar 3 e 4

Menor distância e
agregação pelo mínimo

	1 + 2	3	4	5
1 + 2		12,04	13,00	9,22
3			5,66	10,00
4				14,56
5				

Passo 3: juntar 1+2 e 5

Menor distância e
agregação pelo mínimo

	1 + 2	3 + 4	5
1 + 2		12,04	9,22
3 + 4			10,00
5			



Método *Single* (vizinho mais próximo)

2.i. MÉTODO SINGLE | ANÁLISE DE CLUSTER

38

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1: juntar 1 e 2

Menor distância

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 2: juntar 3 e 4

Menor distância e
agregação pelo mínimo

	1 + 2	3	4	5
1 + 2		12,04	13,00	9,22
3			5,66	10,00
4				14,56
5				

Passo 3: juntar 1+2 e 5

Menor distância e
agregação pelo mínimo

	1 + 2	3 + 4	5
1 + 2		12,04	9,22
3 + 4			10,00
5			

Dado que a menor distância é **9,22**, vamos agrupar as informações dos grupos 1+2 e 5, por meio das **distâncias mínimas** em relação aos demais candidatos:

Distância entre 1+2 e 3+4 = 12,04

Distância entre 5 e 3+4 = 10,00

MÍNIMO é 10,00

Método *Single* (vizinho mais próximo)

2.i. MÉTODO SINGLE | ANÁLISE DE CLUSTER

39

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1: juntar 1 e 2

Menor distância

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 2: juntar 3 e 4

Menor distância e
agregação pelo mínimo

	1 + 2	3	4	5
1 + 2		12,04	13,00	9,22
3			5,66	10,00
4				14,56
5				

Passo 3: juntar 1+2 e 5

Menor distância e
agregação pelo mínimo

	1 + 2	3 + 4	5
1 + 2		12,04	9,22
3 + 4			10,00
5			

Agregação pelo mínimo

	1 + 2 + 5	3 + 4
1 + 2 + 5		10,00
3 + 4		

Método *Single* (vizinho mais próximo)

2.i. MÉTODO SINGLE | ANÁLISE DE CLUSTER

40

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1: juntar 1 e 2

Menor distância

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 2: juntar 3 e 4

Menor distância e agregação pelo mínimo

	1 + 2	3	4	5
1 + 2		12,04	13,00	9,22
3			5,66	10,00
4				14,56
5				

Passo 3: juntar 1+2 e 5

Menor distância e agregação pelo mínimo

	1 + 2	3 + 4	5
1 + 2		12,04	9,22
3 + 4			10,00
5			

Passo 4: juntar 1+2+5 e 3+4

Menor distância e agregação pelo mínimo

	1 + 2 + 5	3 + 4
1 + 2 + 5		10,00
3 + 4		

Método *Single* (vizinho mais próximo)

2.i. MÉTODO SINGLE | ANÁLISE DE CLUSTER

41

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1: juntar 1 e 2

Menor distância

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 2: juntar 3 e 4

Menor distância e agregação pelo mínimo

	1 + 2	3	4	5
1 + 2		12,04	13,00	9,22
3			5,66	10,00
4				14,56
5				

Passo 3: juntar 1+2 e 5

Menor distância e agregação pelo mínimo

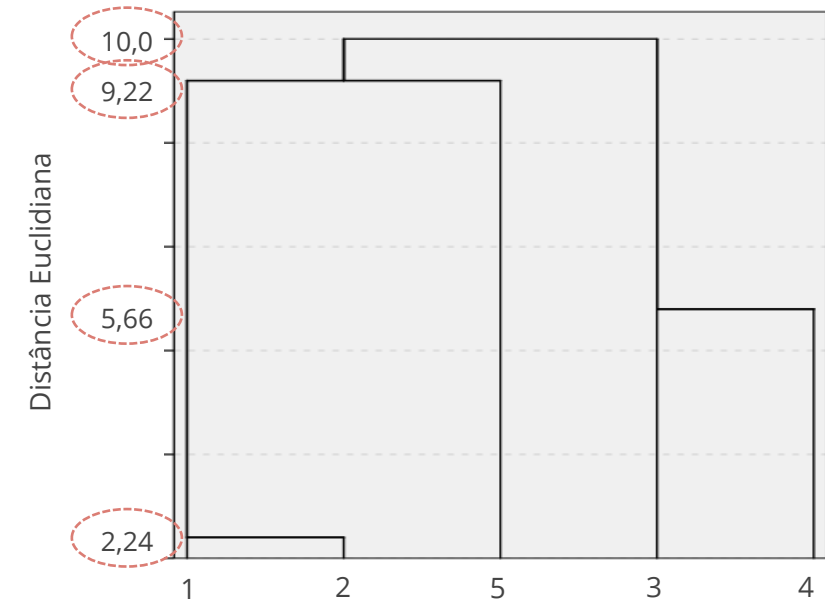
	1 + 2	3 + 4	5
1 + 2		12,04	9,22
3 + 4			10,00
5			

Passo 4: juntar 1+2+5 e 3+4

Menor distância e agregação pelo mínimo

	1 + 2 + 5	3 + 4
1 + 2 + 5		10,00
3 + 4		

Todas as observações em único grupo



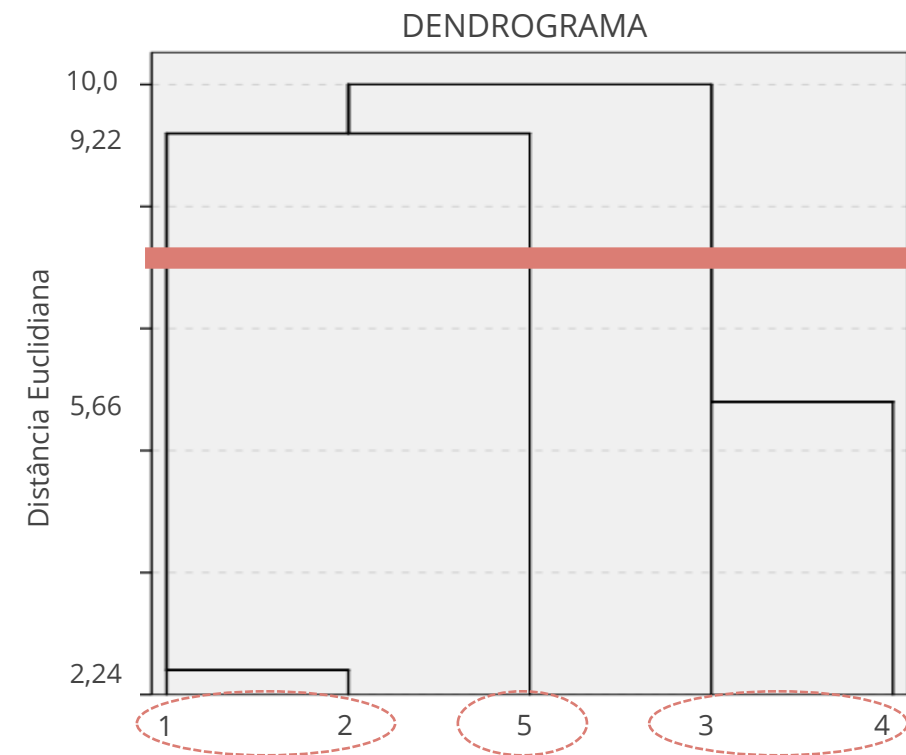
Dendrograma

2. MÉTODO HIERÁRQUICO | ANÁLISE DE CLUSTER

42

O **dendrograma** é uma **representação gráfica** dos passos realizados no agrupamento pelo método hierárquico. Com base na análise do dendrograma é possível investigar o **número de grupos** e **como as observações foram agrupadas**.

Para definir o número de grupos, em geral, observa-se quando o próximo agrupamento é realizado em uma distância muito superior ao agrupamento anterior.



- ✓ O elemento 1 foi agrupado ao 2 na distância 2,24
- ✓ O elemento 3 foi agrupado ao 4 na distância 5,66
- ✓ O grupo (1+2) foi agrupado ao 5 na distância 9,22
- ✓ O grupo (1+2+5) foi agrupado ao grupo (3+4) na distância 10,00

Como a distância entre 9,22 e 5,66 é grande, pode-se sugerir separar os grupos em uma distância superior a 5,657 e inferior a 9,220.

A linha vermelha representa a separação. Abaixo dela, a quantidade de grupos formados; no exemplo, 3 grupos.

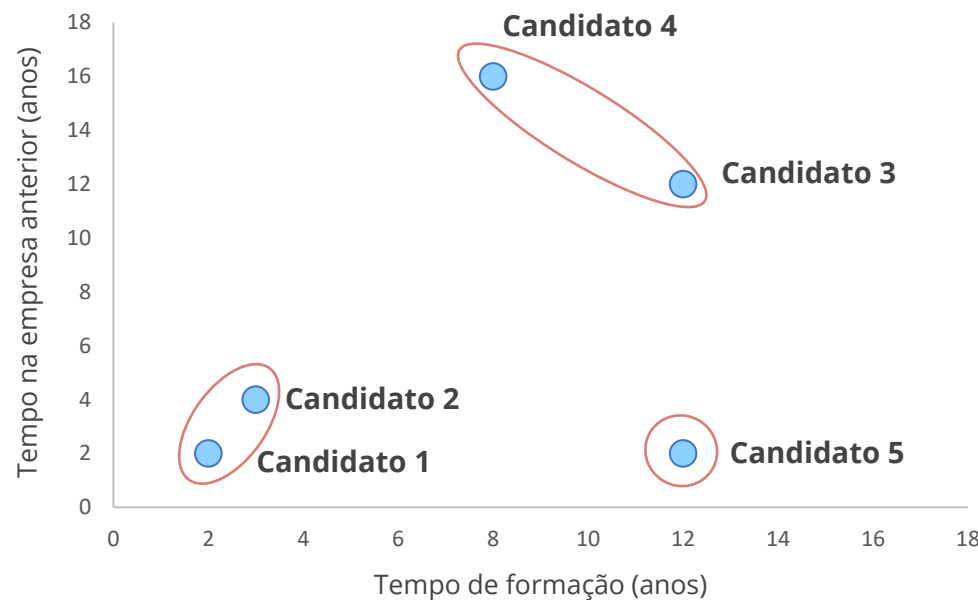
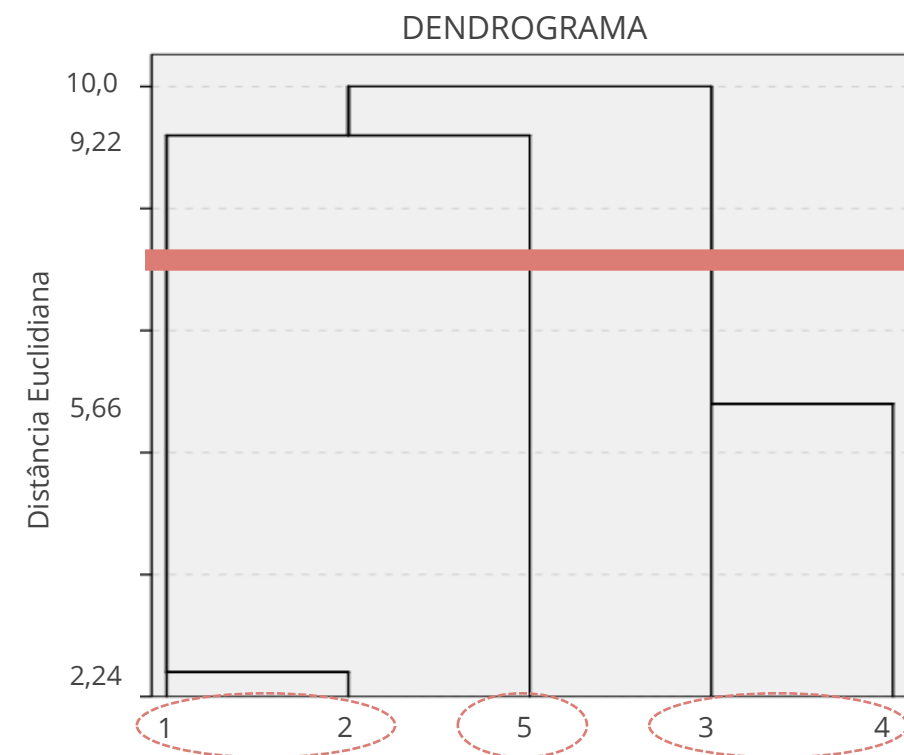
Dendrograma

2. MÉTODO HIERÁRQUICO | ANÁLISE DE CLUSTER

43

O dendrograma (à esquerda) sugere **3 grupos**, assim como observado visualmente, pelo gráfico de dispersão (à direita).

Porém, no caso de 3 ou mais variáveis ou muitas observações, não é possível utilizar o gráfico de dispersão para 'comprovar' a formação de grupos. Por isso o dendrograma é uma representação gráfica muito útil.



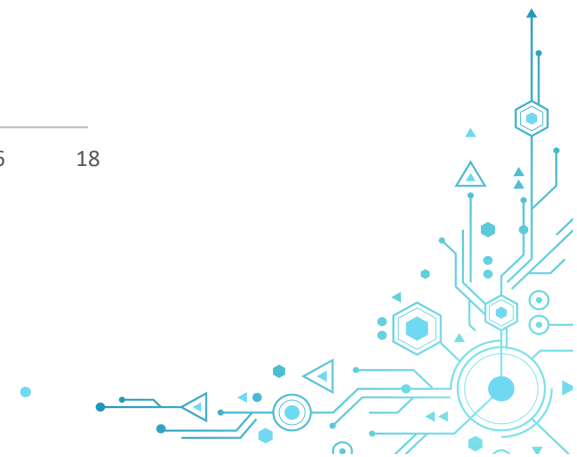
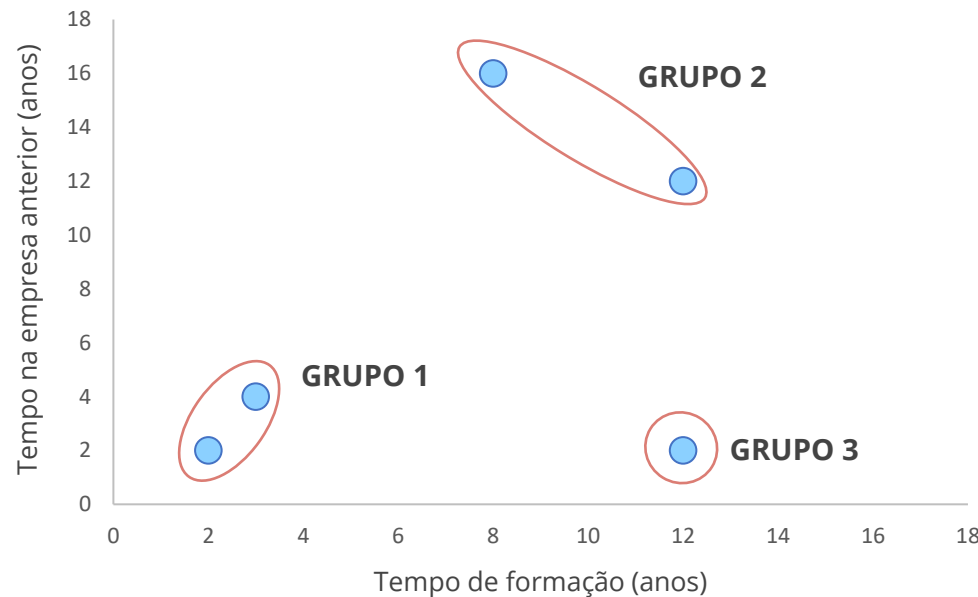
O analista de gestão de pessoas também deseja caracterizar os grupos e criar uma “*persona*” de cada um dos clusters.

Qual a interpretação dos resultados obtidos na Análise de *Cluster*?

O **grupo 1** é formado por candidatos com pouco tempo de formação e pouco tempo na empresa anterior.

O **grupo 2** é formado por candidatos com tempo de formação superior a 7 anos e com tempo na empresa anterior superior a 11 anos.

O **grupo 3** é formado por um candidato com 12 anos de formação e 2 anos na empresa anterior.



Método *Complete* (vizinho mais longe)

2.ii. MÉTODO COMPLETE | ANÁLISE DE CLUSTER

45

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					



Método *Complete* (vizinho mais longe)

2.ii. MÉTODO COMPLETE | ANÁLISE DE CLUSTER

46

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1: juntar 1 e 2

Menor distância

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					



Método *Complete* (vizinho mais longe)

2.ii. MÉTODO COMPLETE | ANÁLISE DE CLUSTER

47

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1: juntar 1 e 2

Menor distância

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Agregamos pelo MÁXIMO

Dado que a menor distância é **2,24**, vamos agrupar as informações dos candidatos 1 e 2, por meio das **distâncias máximas** em relação aos demais candidatos:

Distância entre 1 e 3 = 14,14

Distância entre 2 e 3 = 12,04

MÁXIMO é 14,14

Distância entre 1 e 4 = 15,23

Distância entre 2 e 4 = 13,00

MÁXIMO é 15,23

Distância entre 1 e 5 = 10,00

Distância entre 2 e 5 = 9,22

MÁXIMO é 10,00



Método *Complete* (vizinho mais longe)

2.ii. MÉTODO COMPLETE | ANÁLISE DE CLUSTER

48

Passo 0: matriz de distâncias

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Passo 1: juntar 1 e 2

Menor distância

	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Agregação pelo máximo

	1 + 2	3	4	5
1 + 2		14,14	15,23	10,00
3			5,66	10,00
4				14,56
5				



Discussão entre os métodos

2. MÉTODO HIERÁRQUICO | 2 MÉTODOS

Dado a escolha de uma medida de dissimilaridade (ex.: Distância Euclidiana), precisamos escolher um **critério** para **recalcular a matriz de distâncias**, tais como as duas que foram apresentadas:

- **Single (vizinho mais próximo):** define-se como o MÍNIMO da distância entre um elemento e outro.
- **Complete (vizinho mais longe):** define-se como o MÁXIMO da distância entre um elemento e outro.

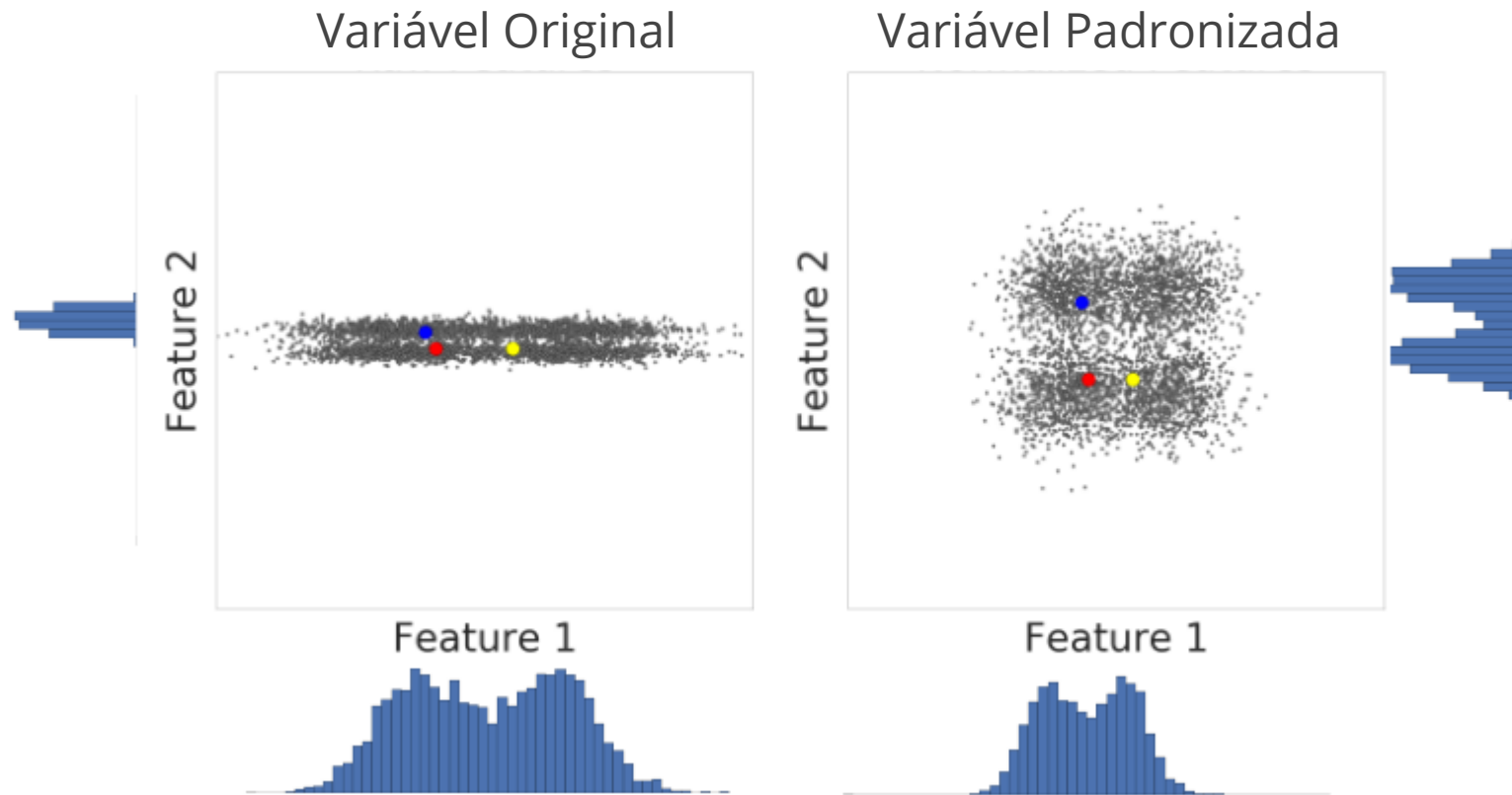


3. Padronização das variáveis



Z-score

3.i. PADRONIZAÇÃO DAS VARIÁVEIS | Z-SCORE



Fonte: <https://developers.google.com/machine-learning/clustering/prepare-data>



- Variáveis com **maior dispersão** (ou seja, maior desvio padrão) têm um **peso maior** no cálculo das distâncias.
- Caso deseje atribuir o mesmo peso para todas as variáveis presentes da análise, é possível utilizar a padronização *Z-score*, que atribui **desvio padrão igual** para todas as variáveis.
- Para se obter uma variável padronizada, deve-se subtrair de cada valor a média e dividir pelo desvio padrão:

$$Z_score = \frac{\text{valor observação} - \text{média}}{\text{desvio padrão}}$$



Case: Hábitos Alimentares

2. MÉTODO HIERÁRQUICO | CASE

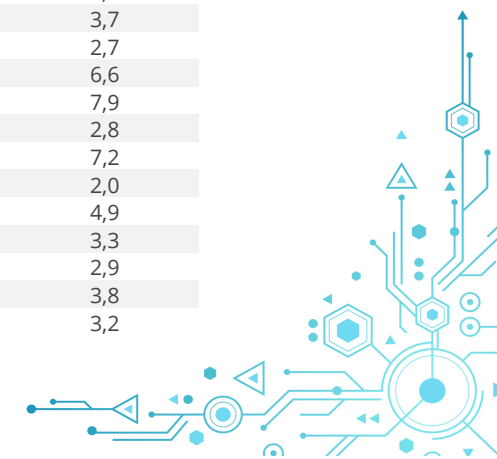
53

Os dados são de uma pesquisa de consumo de alimentos em 25 países da Europa ao longo de determinado período. Nove grupos de alimentos foram analisados: *carne vermelha*, *carne branca*, *ovos*, *leite*, *peixes*, *cereais*, *carboidratos*, *grãos*, *frutas e vegetais*. O objetivo do estudo é agrupar os países segundo comportamentos de hábitos alimentares semelhantes.



Fonte: DASL (*The Data and Story Library*)

Pais	carne_vermelha	carne_branca	ovos	leite	peixes	cereais	carboidratos	graos	fruta_vegetais
Albania	10,1	1,4	0,5	8,9	0,2	42,3	0,6	5,5	1,7
Austria	8,9	14,0	4,3	19,9	2,1	28,0	3,6	1,3	4,3
Belgium	13,5	9,3	4,1	17,5	4,5	26,6	5,7	2,1	4,0
Bulgaria	7,8	6,0	1,6	8,3	1,2	56,7	1,1	3,7	4,2
Czechoslovakia	9,7	11,4	2,8	12,5	2,0	34,3	5,0	1,1	4,0
Denmark	10,6	10,8	3,7	25,0	9,9	21,9	4,8	0,7	2,4
EGermany	8,4	11,6	3,7	11,1	5,4	24,6	6,5	0,8	3,6
Finland	9,5	4,9	2,7	33,7	5,8	26,3	5,1	1,0	1,4
France	18,0	9,9	3,3	19,5	5,7	28,1	4,8	2,4	6,5
Greece	10,2	3,0	2,8	17,6	5,9	41,7	2,2	7,8	6,5
Hungary	5,3	12,4	2,9	9,7	0,3	40,1	4,0	5,4	4,2
Ireland	13,9	10,0	4,7	25,8	2,2	24,0	6,2	1,6	2,9
Italy	9,0	5,1	2,9	13,7	3,4	36,8	2,1	4,3	6,7
Netherlands	9,5	13,6	3,6	23,4	2,5	22,4	4,2	1,8	3,7
Norway	9,4	4,7	2,7	23,3	9,7	23,0	4,6	1,6	2,7
Poland	6,9	10,2	2,7	19,3	3,0	36,1	5,9	2,0	6,6
Portugal	6,2	3,7	1,1	4,9	14,2	27,0	5,9	4,7	7,9
Romania	6,2	6,3	1,5	11,1	1,0	49,6	3,1	5,3	2,8
Spain	7,1	3,4	3,1	8,6	7,0	29,2	5,7	5,9	7,2
Sweden	9,9	7,8	3,5	24,7	7,5	19,5	3,7	1,4	2,0
Switzerland	13,1	10,1	3,1	23,8	2,3	25,6	2,8	2,4	4,9
UK	17,4	5,7	4,7	20,6	4,3	24,3	4,7	3,4	3,3
USSR	9,3	4,6	2,1	16,6	3,0	43,6	6,4	3,4	2,9
WGermany	11,4	12,5	4,1	18,8	3,4	18,6	5,2	1,5	3,8
Yugoslavia	4,4	5,0	1,2	9,5	0,6	55,9	3,0	5,7	3,2



Case: Hábitos Alimentares

2. MÉTODO HIERÁRQUICO | CASE

54

Os dados são de uma pesquisa de consumo de alimentos em 25 países da Europa ao longo de determinado período. Nove grupos de alimentos foram analisados: *carne vermelha*, *carne branca*, *ovos*, *leite*, *peixes*, *cereais*, *carboidratos*, *grãos*, *frutas e vegetais*. O objetivo do estudo é agrupar os países segundo comportamentos de hábitos alimentares semelhantes.



Fonte: DASL (*The Data and Story Library*)

Variável	Descrição
carne_vermelha	Índice de consumo de carne vermelha (em toneladas)
carne_branca	Índice de consumo de carne branca (em toneladas)
ovos	Índice de consumo de ovos (em milhões)
leite	Índice de consumo de leite (em milhões de litros)
peixes	Índice de consumo de peixes (em toneladas)
cereais	Índice de consumo de cereais (em toneladas)
carboidratos	Índice de consumo de carboidratos (em toneladas)
graos	Índice de consumo de grãos (em toneladas)
fruta_vegetais	Índice de consumo de frutas e vegetais (em toneladas)

Arquivo: Consumo_Alimentos.xlsx



Case: Hábitos Alimentares

2. MÉTODO HIERÁRQUICO | CASE

55

Os dados são de uma pesquisa de consumo de alimentos em 25 países da Europa ao longo de determinado período. Nove grupos de alimentos foram analisados: *carne vermelha*, *carne branca*, *ovos*, *leite*, *peixes*, *cereais*, *carboidratos*, *grãos*, *frutas e vegetais*. O objetivo do estudo é agrupar os países segundo comportamentos de hábitos alimentares semelhantes.



Fonte: DASL (*The Data and Story Library*)

- (a) Faça uma análise exploratória da base de dados (obtenha as medidas de posição e dispersão).
- (b) Para as variáveis *leite* e *carboidratos*, comente os quartis: Q1, Q2 (mediana) e Q3.
- (c) Considerando o histograma das variáveis *leite* e *carboidratos*, as distribuições são simétricas?
- (d) Considerando as variáveis *carne vermelha* e *carne branca*, qual possui a maior variabilidade?
- (e) Existem outliers nas variáveis *carne vermelha* e *carne branca*?
- (f) Padronize as variáveis.
- (g) Calcule a matriz de distâncias euclidianas entre os 25 países.
- (h) Faça a análise de agrupamento com as variáveis padronizadas, usando os 2 métodos apresentados. Escolha um dos métodos e justifique a quantidade de grupos, após a análise do dendrograma.
- (i) Pelo dendrograma do método *Complete*, qual país é mais semelhante à Romênia?
- (j) Analise as características de cada grupo, a partir dos box plots. Comente os resultados.



Arquivo: Consumo_Alimentos.xlsx



Case: Varejo

2. MÉTODO HIERÁRQUICO | CASE

56

Uma empresa de e-commerce deseja agrupar seus clientes para criar diferentes ações de marketing, com base em três variáveis: quantidade de compras, valor dos produtos e nota de revisão.



customer_id	qtde_compras	valor_produtos	nota_revisao
03d01c3308507d5d861af0d89b65beee	1	33,0	4
70760f5ca54f7826fbc14b679eb949bd	1	80,0	5
43cd92deaa3d542fa32fdf4ca3089f51	6	240,0	2
676ea4c495818f6654dd38d006cfb1d7	6	86,5	5
8702a62684cd9a0ad5a391017c6939d6	4	166,7	2
b154a09d611816a2bd59fa2582e5beeb8	3	25,0	5
efd22fcffe47d73526c48448b1c47292	4	149,9	10
2de7c1adffd1bb406b6bd0b76ddfe85e	1	179,9	4
a2f0e5f633e77e713c7e49a836876c78	2	45,0	6
...

Arquivo: Varejo.xlsx



Case: Varejo

2. MÉTODO HIERÁRQUICO | CASE

57

Uma empresa de e-commerce deseja agrupar seus clientes para criar diferentes ações de marketing, com base em três variáveis: quantidade de compras, valor dos produtos e nota de revisão.



Variável	Descrição
customer_id	Código do cliente
qtde_compras	Quantidade de compras do cliente no último ano
valor_produtos	Valor médio dos produtos comprados no último ano
nota_revisao	Nota média da revisão dos produtos comprados no último ano

Arquivo: Varejo.xlsx



Case: Varejo

2. MÉTODO HIERÁRQUICO | CASE

58

Uma empresa de e-commerce deseja agrupar seus clientes para criar diferentes ações de marketing, com base em três variáveis: quantidade de compras, valor dos produtos e nota de revisão.



- (a) Faça uma análise exploratória da base de dados (obtenha as medidas de posição e dispersão).
- (b) Para todas as variáveis, comente os quartis: Q1, Q2 (mediana) e Q3.
- (c) Considerando o histograma das variáveis, as distribuições são simétricas?
- (d) Existem outliers nas variáveis?
- (e) Padronize as variáveis.
- (f) Calcule a matriz de distâncias euclidianas.
- (g) Faça a análise de agrupamento com as variáveis padronizadas, usando os 2 métodos apresentados. Escolha um dos métodos e justifique a quantidade de grupos, após a análise do dendrograma.
- (h) Analise as características de cada grupo, a partir dos box plots. Comente os resultados.
- (i) Crie um plano de ação de e-mail marketing de um ano de cadastro para cada grupo.

Arquivo: Varejo.xlsx



Case: Imobiliário

2. MÉTODO HIERÁRQUICO | CASE

59

Uma imobiliária deseja agrupar seus imóveis à venda para atribuí-los a diferentes corretores. Os imóveis serão agrupados por sua idade, distância ao metrô, comércios próximos e valor por m².



Id_Imovel	Idade_imovel	Distancia_metro_Km	Comercios_proximos	Mil_reais_m2
1	32,0	1,08	10	7,58
2	19,5	1,40	9	8,44
3	13,3	1,54	5	9,46
4	13,3	1,54	5	10,96
5	5,0	1,46	5	8,62
6	7,1	1,87	3	6,42
7	34,5	1,57	7	8,06
8	20,3	1,38	6	9,34
9	31,7	2,10	1	3,76
...

Arquivo: Imobiliario.xlsx



Case: Imobiliário

2. MÉTODO HIERÁRQUICO | CASE

Uma imobiliária deseja agrupar seus imóveis à venda para atribuí-los a diferentes corretores. Os imóveis serão agrupados por sua idade, distância ao metrô, comércios próximos e valor por m².



Variável	Descrição
Id_Imovel	ID do imóvel
Idade_imovel	Idade do imóvel
Distancia_metro_Km	Distância ao metrô mais próximo, em km
Comercios_proximos	Quantidade média de comércios próximos
Mil_reais_m2	Valor do imóvel por m², em milhares de reais

Arquivo: Imobiliario.xlsx



Case: Imobiliário

2. MÉTODO HIERÁRQUICO | CASE

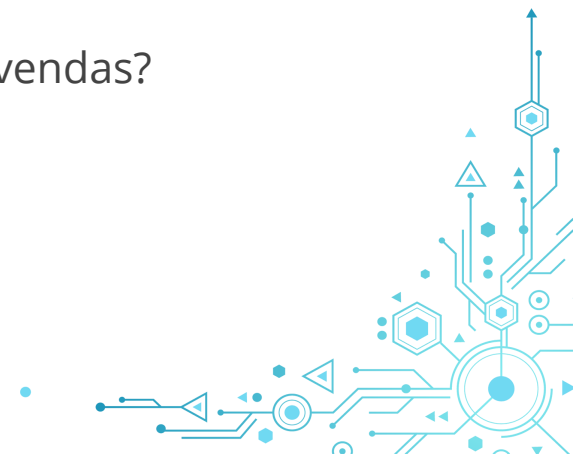
61

Uma imobiliária deseja agrupar seus imóveis à venda para atribuí-los a diferentes corretores. Os imóveis serão agrupados por sua idade, distância ao metrô, comércios próximos e valor por m².



- (a) Faça uma análise exploratória da base de dados (obtenha as medidas de posição e dispersão).
- (b) Para todas as variáveis, comente os quartis: Q1, Q2 (mediana) e Q3.
- (c) Considerando o histograma das variáveis, as distribuições são simétricas?
- (d) Existe outlier nas variáveis?
- (e) Padronize as variáveis.
- (f) Calcule a matriz de distâncias euclidianas.
- (g) Faça a análise de agrupamento com as variáveis padronizadas, usando os 2 métodos apresentados. Escolha um dos métodos e justifique a quantidade de grupos, após a análise do dendrograma.
- (h) Analise as características de cada grupo, a partir dos box plots. Comente os resultados.
- (i) Para qual grupo você atribuiria seus corretores mais experientes, com maior capacidade de vendas?

Arquivo: Imobiliario.xlsx



Case: Municípios

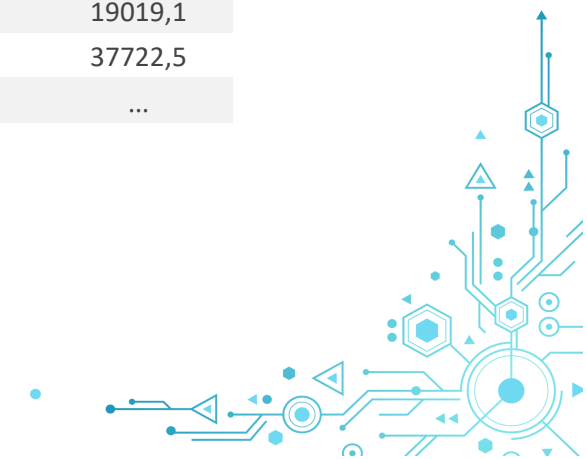
2. MÉTODO HIERÁRQUICO | 2 MÉTODOS

62

Os dados são relacionados a indicadores econômicos de cada município do estado de São Paulo. O objetivo do estudo é agrupar estes municípios segundo o comportamento das variáveis econômicas.



Município	pop15	pop60	hab	area	taxa	esgoto	emprego	pib
Guarulhos	22,5	9,6	3830,3	318,7	16,9	86,9	29,0	44670,7
Campinas	18,2	13,8	1358,4	794,4	13,9	87,0	40,5	42766,0
São Bernardo do Campo	19,4	11,9	1868,0	409,5	14,1	90,3	38,2	34185,3
Santo André	18,0	15,0	3846,7	175,8	13,1	94,5	31,8	18085,1
Osasco	20,9	11,1	10263,6	65,0	15,7	83,8	26,3	39198,9
São José dos Campos	20,4	11,6	572,2	1099,4	14,6	93,3	33,6	28089,1
Ribeirão Preto	18,2	13,8	927,5	651,0	13,2	97,5	38,2	20300,8
Sorocaba	19,4	12,2	1302,3	449,8	15,0	97,8	35,2	19019,1
Santos	16,3	20,3	1494,2	280,7	11,7	95,3	45,2	37722,5
...



Case: Municípios

2. MÉTODO HIERÁRQUICO | 2 MÉTODOS

63

Os dados são relacionados a indicadores econômicos de cada município do estado de São Paulo. O objetivo do estudo é agrupar estes municípios segundo o comportamento das variáveis econômicas.



Variável	Descrição
pop15	% de habitantes com até 15 anos de idade
pop60	% de habitantes com 60 anos de idade ou mais
hab	Quantidade de habitantes por km ² (densidade demográfica)
area	Área, em km ²
taxa	Taxa de natalidade
esgoto	% de domicílios com acesso a saneamento básico
emprego	% de habitantes com emprego formal
pib	PIB per capita

Arquivo: Municipios.xlsx

@2021 LABDATA FIA. Copyright all rights reserved.



Case: Municípios

2. MÉTODO HIERÁRQUICO | 2 MÉTODOS

64

Os dados são relacionados a indicadores econômicos de cada município do estado de São Paulo. O objetivo do estudo é agrupar estes municípios segundo o comportamento das variáveis econômicas.



- (a) Faça uma análise exploratória da base de dados (obtenha as medidas de posição e dispersão).
- (b) Considerando o histograma das variáveis hab e pib, as distribuições são simétricas?
- (c) Considerando as variáveis pop15 e pop60, qual possui a maior variabilidade?
- (d) Existe outlier nas variáveis pop15 e pop60?
- (e) Padronize as variáveis.
- (f) Calcule a matriz de distâncias euclidianas.
- (g) Faça a análise de agrupamento com as variáveis padronizadas, usando os 2 métodos apresentados. Quantos grupos poderiam ser sugeridos?

Arquivo: Municipios.xlsx



Case: Municípios

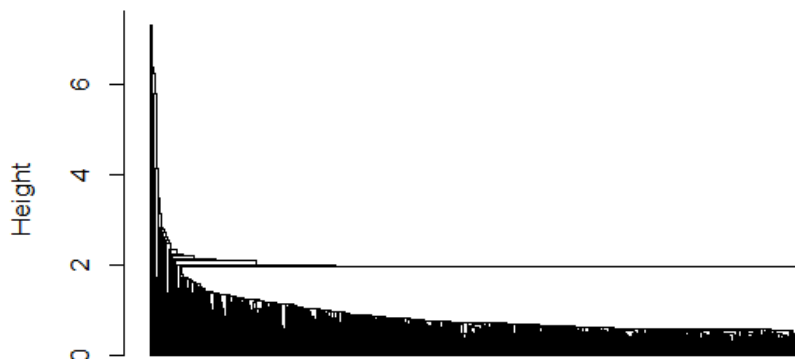
2. MÉTODO HIERÁRQUICO | 2 MÉTODOS

65

Os dados são relacionados a indicadores econômicos de cada município do estado de São Paulo. O objetivo do estudo é agrupar estes municípios segundo o comportamento das variáveis econômicas.

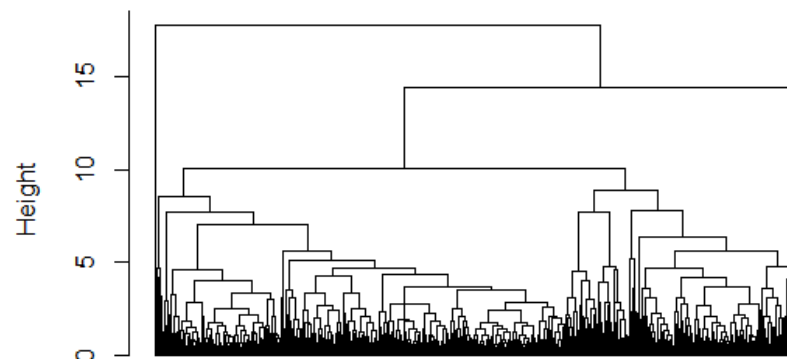


Metodo Single



distancia
hclust (*, "single")

Metodo Complete



distancia
hclust (*, "complete")

Arquivo: Municipios.xlsx

@2021 LABDATA FIA. Copyright all rights reserved.



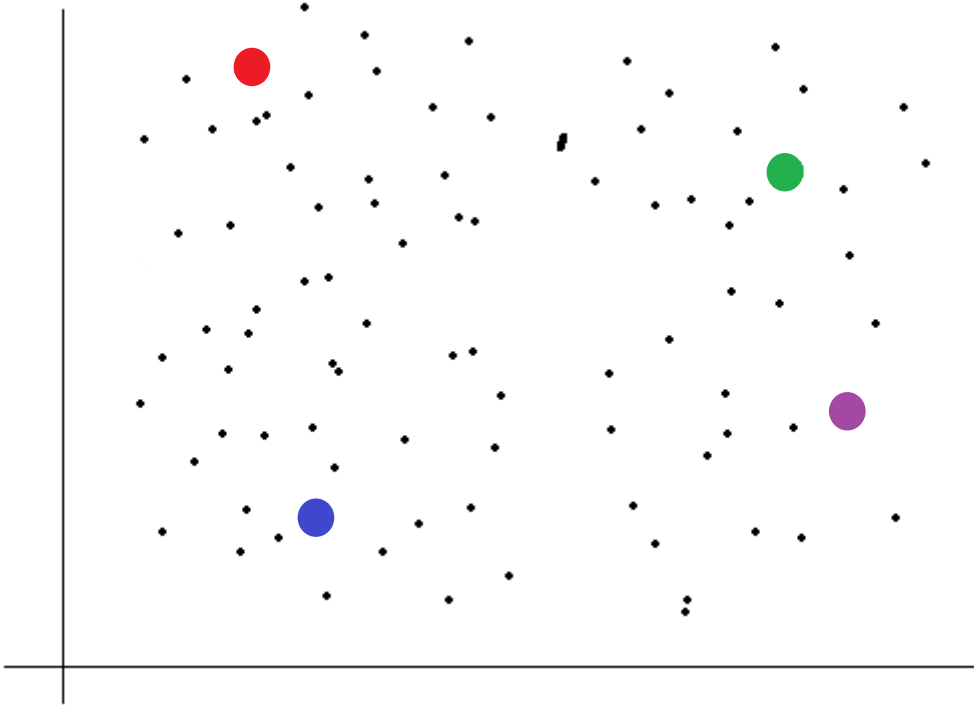
4. Método de Partição: K-médias



Processo de Agrupamento

4. MÉTODO DE PARTIÇÃO K-MÉDIAS | ANÁLISE DE CLUSTER

67



O algoritmo cria **k centroides** (sementes) aleatórios.

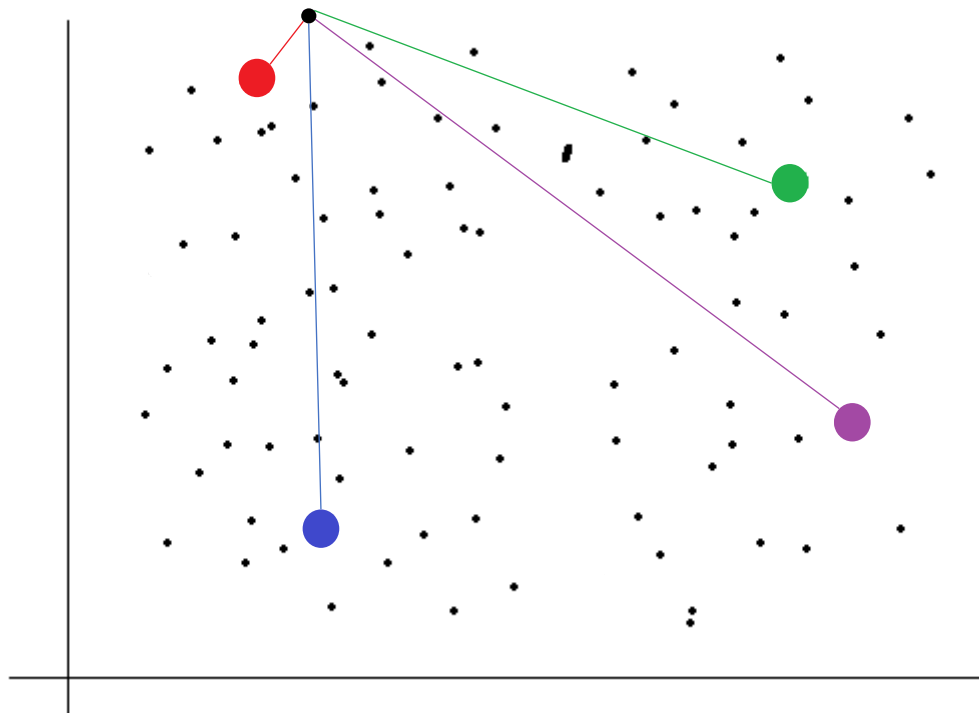
Neste caso, $k = 4$.



Processo de Agrupamento

4. MÉTODO DE PARTIÇÃO K-MÉDIAS | ANÁLISE DE CLUSTER

68



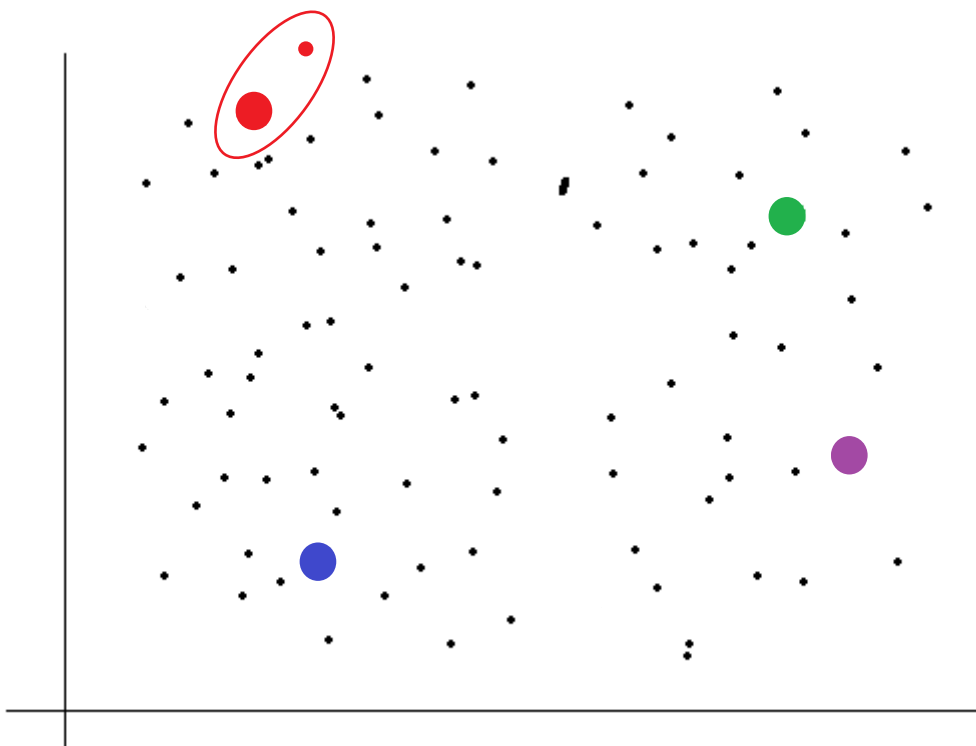
Calcula-se a Distância Euclidiana de 1 observação para os 4 centroides. Esta observação está mais próxima de qual centroide?



Processo de Agrupamento

4. MÉTODO DE PARTIÇÃO K-MÉDIAS | ANÁLISE DE CLUSTER

69



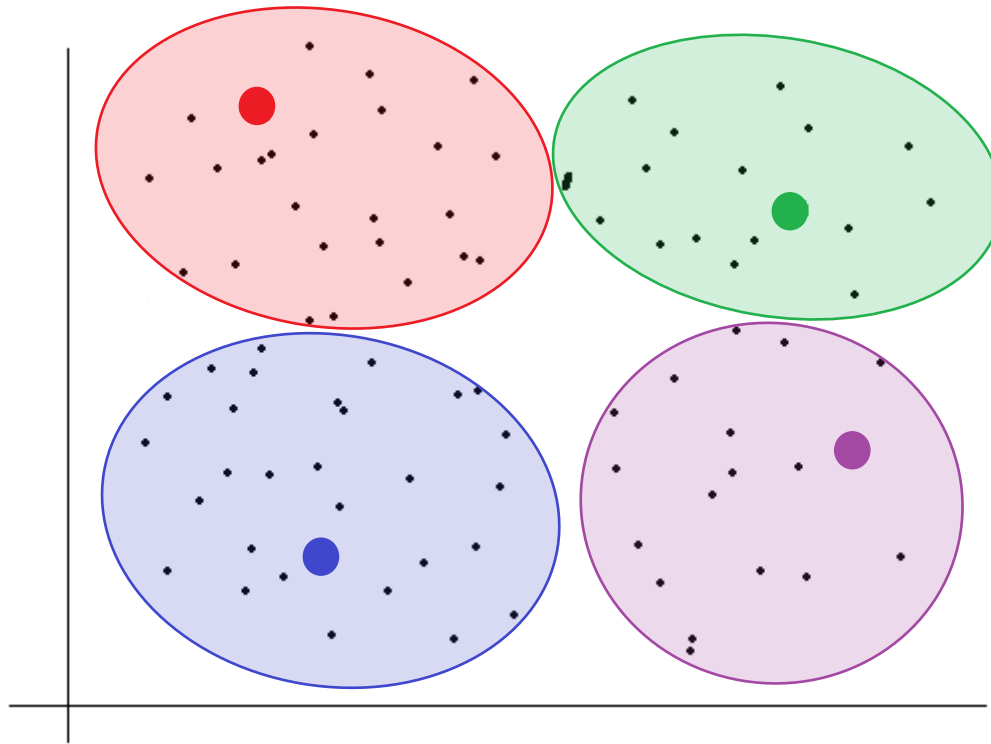
Por estar mais próxima do centroide vermelho, esta observação passa a pertencer ao **grupo vermelho**.



Processo de Agrupamento

4. MÉTODO DE PARTIÇÃO K-MÉDIAS | ANÁLISE DE CLUSTER

70



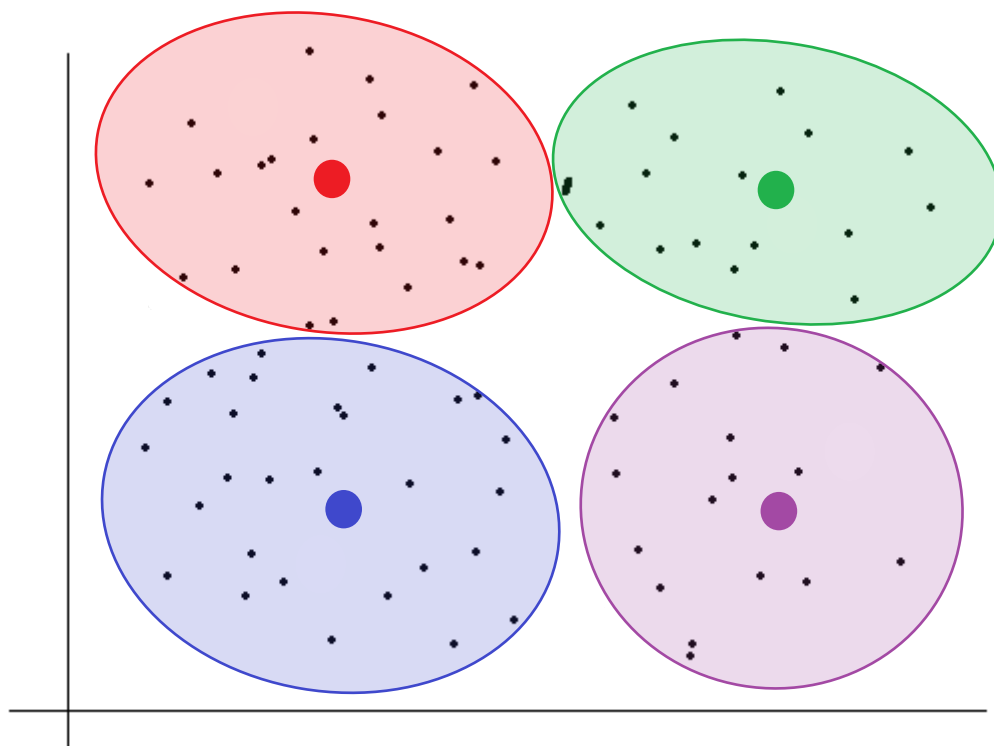
Repete-se o processo para todas as observações, até que se finalizem os grupos.



Processo de Agrupamento

4. MÉTODO DE PARTIÇÃO K-MÉDIAS | ANÁLISE DE CLUSTER

71



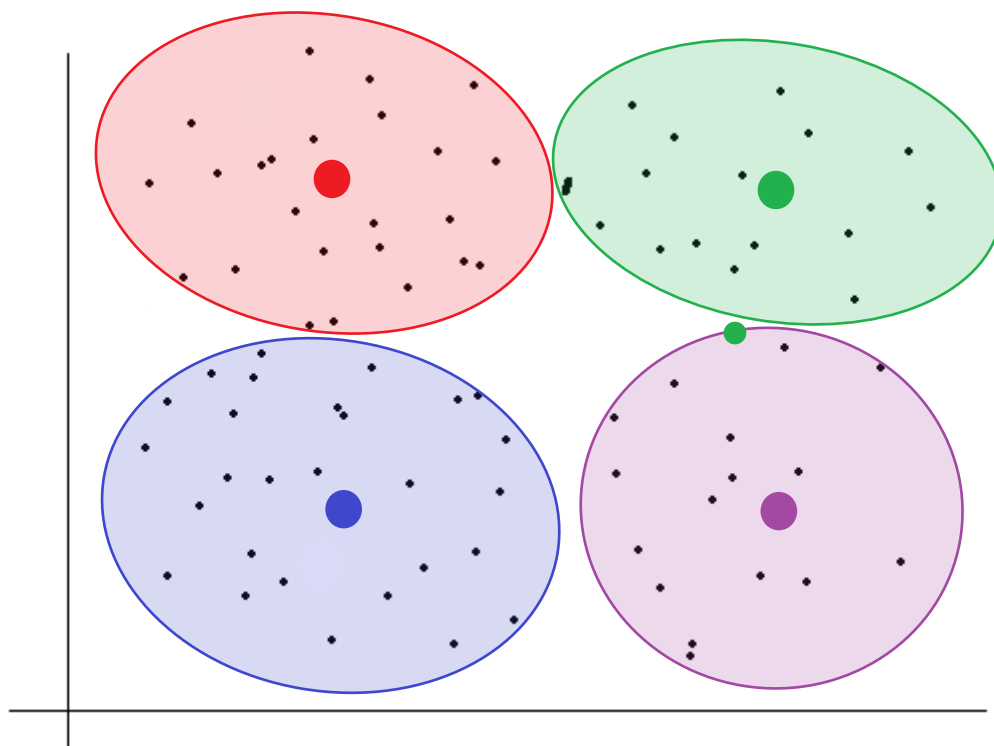
Então, os centroides (“centros”) de cada grupo são **atualizados** com base nas coordenadas de seus elementos.



Processo de Agrupamento

4. MÉTODO DE PARTIÇÃO K-MÉDIAS | ANÁLISE DE CLUSTER

72



Os passos anteriores são realizados de forma iterativa, a fim de **reclassificar observações** que estejam mais próximas do centróide de outro grupo.

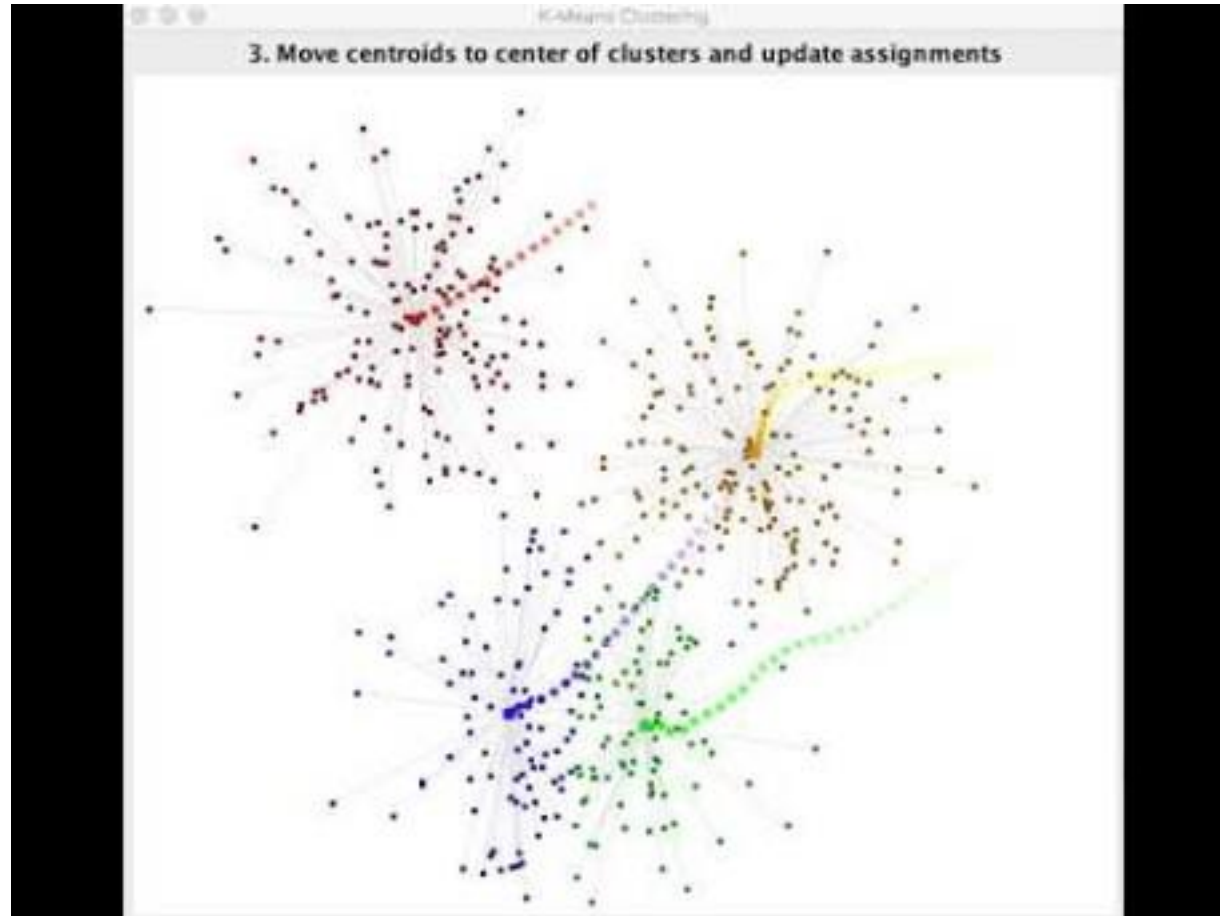
O algoritmo é finalizado quando se atinge convergência; ou seja, quando nenhuma observação muda mais de grupo.



Processo Iterativo

4. MÉTODO DE PARTIÇÃO K-MÉDIAS | ANÁLISE DE CLUSTER

73



Exemplo do processo iterativo descrito anteriormente.

<https://www.youtube.com/watch?v=nXY6PxAaOk0>



Utiliza um procedimento de aproximação. Por isso, pode ser usado em grandes bancos de dados.

Considerações sobre o método

- O número de clusters (k) precisa ser previamente definido.
- As coordenadas do centroide de cada grupo são definidas como a média entre as coordenadas de seus elementos.
- A cada passo, os elementos são agrupados no *cluster* com o centroide mais próximo, com subsequentes recálculos dos centros.
- As observações são agrupadas nos centroides até que as partições encontradas satisfaçam algum critério de qualidade especificado.



Case: Municípios

4. MÉTODO DE PARTIÇÃO K-MÉDIAS | CASE

75

Os dados são relacionados a indicadores econômicos de cada município do estado de São Paulo. O objetivo do estudo é agrupar estes municípios segundo o comportamento das variáveis econômicas.



Com a mesma base de dados do exercício anterior:

- (a) Utilize o método K-médias para 2, 3, 4 e 5 grupos. Qual número de grupos é melhor?
- (b) Caracterize os grupos.
- (c) Uma varejista que só possui lojas na capital (São Paulo) deseja escolher outras cidades para criar mais lojas. Qual grupo de cidades você escolheria?

Arquivo: Municipios.xlsx

@2021 LABDATA FIA. Copyright all rights reserved.

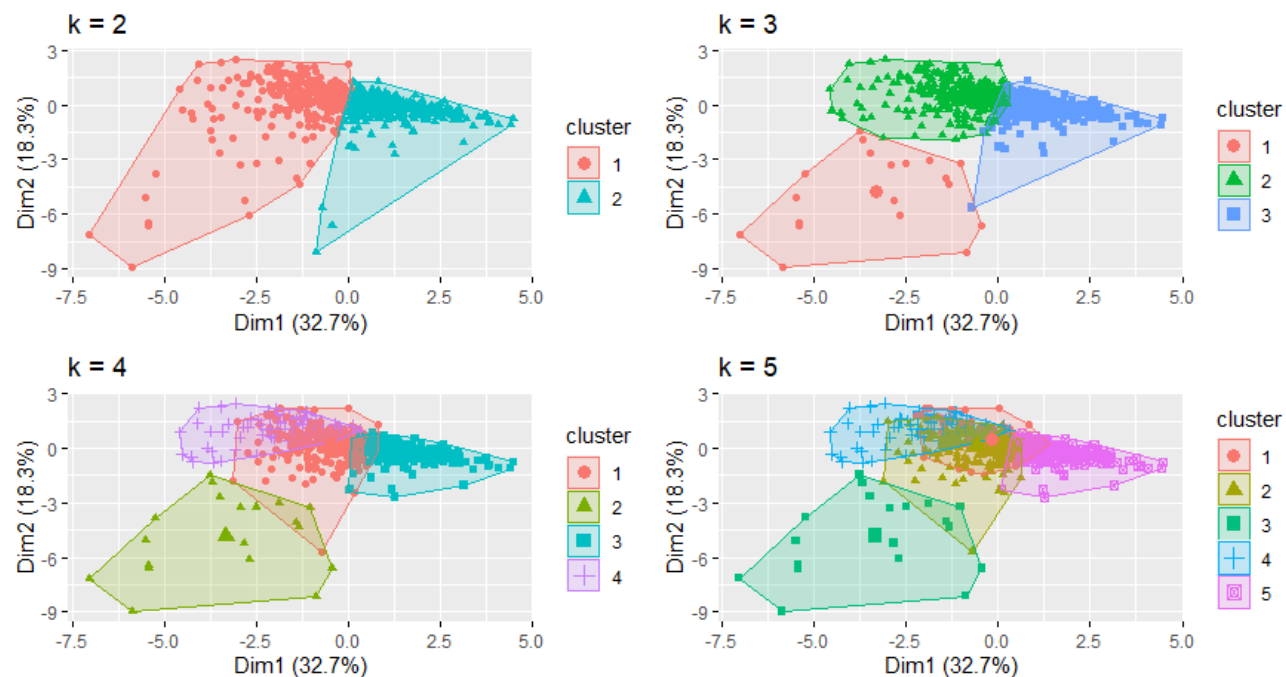


Case: Municípios

4. MÉTODO DE PARTIÇÃO K-MÉDIAS | CASE

76

Os dados são relacionados a indicadores econômicos de cada município do estado de São Paulo. O objetivo do estudo é agrupar estes municípios segundo o comportamento das variáveis econômicas.



Arquivo: Municipios.xlsx

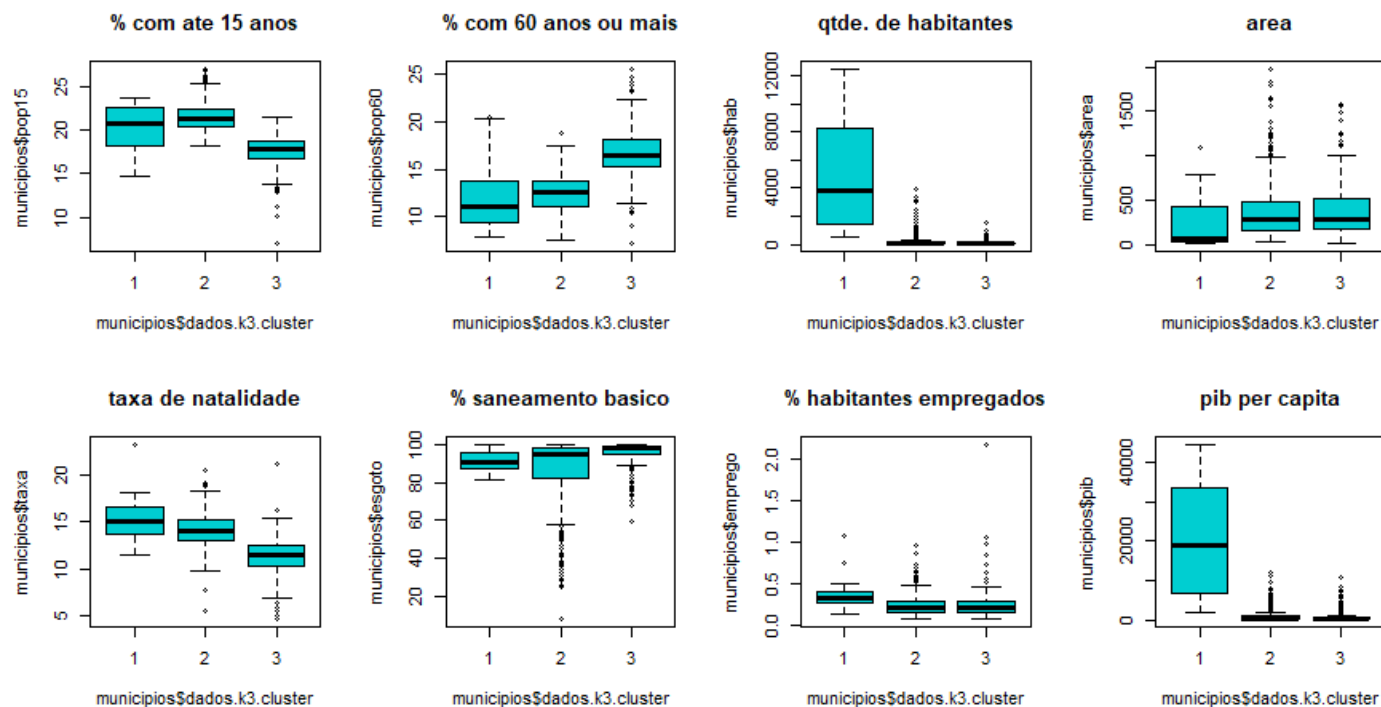


Case: Municípios

4. MÉTODO DE PARTIÇÃO K-MÉDIAS | CASE

77

Os dados são relacionados a indicadores econômicos de cada município do estado de São Paulo. O objetivo do estudo é agrupar estes municípios segundo o comportamento das variáveis econômicas.



Arquivo: Municipios.xlsx



Case: Limite de cartão de crédito

4. MÉTODO DE PARTIÇÃO K-MÉDIAS | CASE

78

Os dados são relacionados a indicadores financeiros de clientes de cartão de crédito. O objetivo do estudo é agrupar os clientes para criar um plano de ação de negócio para cada grupo, dadas as suas características. Este plano de ação será executado por meio de um envio de e-mail marketing.



CLIENTE	LIMITE	IDADE	PERC_USO_CARTAO
1	2000	24	0,20
2	12000	26	0,02
3	9000	34	0,32
4	5000	37	0,94
5	5000	57	0,17
6	5000	37	1,29
7	50000	29	0,74
8	10000	23	0,12
9	14000	28	0,08
...

Arquivo: Limite.xlsx



Case: Limite de cartão de crédito

4. MÉTODO DE PARTIÇÃO K-MÉDIAS | CASE

79

Os dados são relacionados a indicadores financeiros de clientes de cartão de crédito. O objetivo do estudo é agrupar os clientes para criar um plano de ação de negócio para cada grupo, dadas as suas características. Este plano de ação será executado por meio de um envio de e-mail marketing.



Variável	Descrição
CLIENTE	ID do cliente
LIMITE	Limite total do cartão solicitado pelo cliente
IDADE	Idade do cliente
PERC_USO_CARTAO	Percentual médio histórico de uso de limite do cliente

Arquivo: Limite.xlsx



Case: Limite de cartão de crédito

4. MÉTODO DE PARTIÇÃO K-MÉDIAS | CASE

80

Os dados são relacionados a indicadores financeiros de clientes de cartão de crédito. O objetivo do estudo é agrupar os clientes para criar um plano de ação de negócio para cada grupo, dadas as suas características. Este plano de ação será executado por meio de um envio de e-mail marketing.



- (a) Faça a análise exploratória de cada variável individualmente.
- (b) Interprete a média, mediana, Q1, Q2 e Q3 para todas as variáveis.
- (c) Faça o box plot e histograma para todas as variáveis.
- (d) Padronize as variáveis.
- (e) Utilize o método K-médias para 2, 3, 4 e 5 grupos. Qual número de grupos é melhor?
- (f) Caracterize os grupos.
- (g) Crie um plano de ação para e-mail marketing para cada grupo.

Arquivo: Limite.xlsx

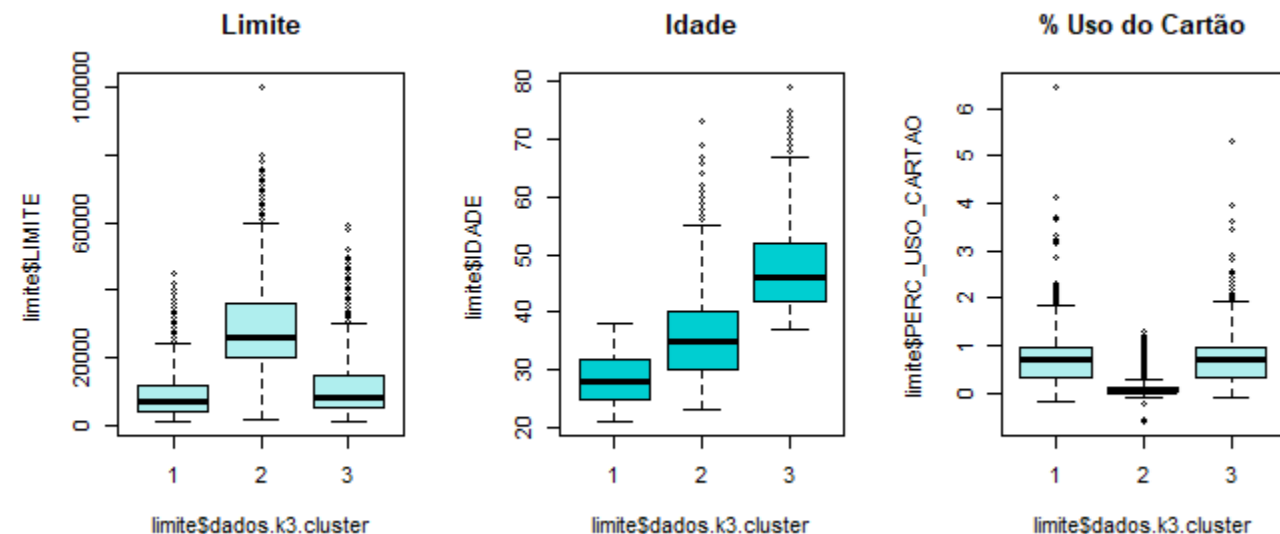


Case: Limite de cartão de crédito

4. MÉTODO DE PARTIÇÃO K-MÉDIAS | CASE

81

Os dados são relacionados a indicadores financeiros de clientes de cartão de crédito. O objetivo do estudo é agrupar os clientes para criar um plano de ação de negócio para cada grupo, dadas as suas características. Este plano de ação será executado por meio de um envio de e-mail marketing.



1. GRUPO 1: Jovens

O grupo 1 apresenta limite mais baixo, menor idade e percentual de uso do cartão alto. Para este grupo, pode-se enviar um e-mail marketing ressaltando as funcionalidades do aplicativo do cartão.

2. GRUPO 2: Low-Users

O grupo 2 apresenta alto limite e baixo percentual de uso do cartão. Pode-se ressaltar os benefícios do uso do cartão, como programa de pontos, para aumentar sua utilização.

3. GRUPO 3: Experientes

O grupo 3 apresenta limite mais baixo, maior idade e percentual de uso do cartão alto. Para este grupo, pode-se enviar um e-mail marketing ressaltando a segurança financeira da empresa e o contato do gerente do cartão de crédito.

Arquivo: Limite.xlsx

5. Exercícios



Case: Serviço de Entregas

5. EXERCÍCIOS

83

Uma empresa de serviço de delivery de comida pronta tem o objetivo de fazer ações de relacionamento e reconhecimento com seus clientes. Para isso, gostaria de identificar os perfis de clientes com base na frequência de pedidos, valor desses pedidos, distância entre o estabelecimento e a residência do cliente, e o tempo de entrega.



Fonte: <https://www.kaggle.com/asaumya/k-means-clustering-food-delivery-case-study/data>

ID_Cliente	N_pedidos	Valor	Dist_m_restaurante	Tempo_m_entrega
28	1	558	0,8	33
122	2	2685	2,6	24
152	12	2608	2,6	31
173	9	1634	1,2	45
204	2	562	2	53
397	13	3045	3,7	38
507	3	379	0,8	39
784	26	4782	2,7	44
931	3	2422	1,7	26
...

Arquivo: Serviço_Entregas.xlsx

@2021 LABDATA FIA. Copyright all rights reserved.



Case: Serviço de Entregas

5. EXERCÍCIOS

84

Uma empresa de serviço de delivery de comida pronta tem o objetivo de fazer ações de relacionamento e reconhecimento com seus clientes. Para isso, gostaria de identificar os perfis de clientes com base na frequência de pedidos, valor desses pedidos, distância entre o estabelecimento e a residência do cliente, e o tempo de entrega.

Fonte: <https://www.kaggle.com/asaumya/k-means-clustering-food-delivery-case-study/data>



Variável	Descrição
ID_Cliente	ID do Cliente
N_pedidos	Número de pedidos totais
Valor	Valor total dos pedidos (R\$)
Dist_m_restaurante	Distância média entre o(s) restaurante(s) e o cliente (km)
Tempo_m_entrega	Tempo médio dos pedidos

Arquivo: Serviço_Entregas.xlsx



Case: Serviço de Entregas

5. EXERCÍCIOS

85

Uma empresa de serviço de delivery de comida pronta tem o objetivo de fazer ações de relacionamento e reconhecimento com seus clientes. Para isso, gostaria de identificar os perfis de clientes com base na frequência de pedidos, valor desses pedidos, distância entre o estabelecimento e a residência do cliente, e o tempo de entrega.

Fonte: <https://www.kaggle.com/asaumya/k-means-clustering-food-delivery-case-study/data>



- (a) Faça a análise exploratória de cada variável individualmente.
- (b) Interprete a média, mediana, Q1, Q2 e Q3 para todas as variáveis.
- (c) Faça o box plot e histograma para todas as variáveis.
- (d) Padronize as variáveis.
- (e) Utilize o método K-médias para 2, 3, 4 e 5 grupos. Qual número de grupos é melhor?
- (f) Caracterize os grupos.
- (g) Crie um plano de ações de marketing para cada grupo.

Arquivo: Serviço_Entregas.xlsx



Case: Marketing Cartão

5. EXERCÍCIOS

86

Uma instituição financeira, emissora de cartões de crédito, deseja identificar os diferentes perfis transacionais em relação ao uso do cartão de crédito, para trabalhar em ações de marketing e comunicação de forma segmentada. O conjunto de dados resume o comportamento de uso de cerca de 8.950 titulares de cartão de crédito, ativos durante os últimos 6 meses. O arquivo está estruturado na visão cliente.

Fonte: <https://www.kaggle.com/mirichoi0218/insurance>



ID_CLIENTE	LIMITE_DISP	VALOR_ENTRADA	QTDE_COMPRAS
C10001	40,9	0,0	2
C10002	3202,5	6442,9	0
C10003	2495,1	0,0	12
C10004	1666,7	205,8	1
C10005	817,7	0,0	1
C10006	1809,8	0,0	8
C10007	627,3	0,0	64
C10008	1823,7	0,0	12
C10009	1014,9	0,0	5
...

Arquivo: Marketing_Cartao.xlsx



Case: Marketing Cartão

5. EXERCÍCIOS

87

Uma instituição financeira, emissora de cartões de crédito, deseja identificar os diferentes perfis transacionais em relação ao uso do cartão de crédito, para trabalhar em ações de marketing e comunicação de forma segmentada. O conjunto de dados resume o comportamento de uso de cerca de 8.950 titulares de cartão de crédito, ativos durante os últimos 6 meses. O arquivo está estruturado na visão cliente.



Fonte: <https://www.kaggle.com/mirichoi0218/insurance>

Variável	Descrição
ID_CLIENTE	Identificação do titular do cartão de crédito
LIMITE_DISP	Valor do limite restante em seu cartão para fazer compras
VALOR_ENTRADA	Adiantamento em dinheiro pago pelo cliente
QTDE_COMPRAS	Quantidade de compras no total

Arquivo: Marketing_Cartao.xlsx



Case: Marketing Cartão

5. EXERCÍCIOS

88

Uma instituição financeira, emissora de cartões de crédito, deseja identificar os diferentes perfis transacionais em relação ao uso do cartão de crédito, para trabalhar em ações de marketing e comunicação de forma segmentada. O conjunto de dados resume o comportamento de uso de cerca de 8.950 titulares de cartão de crédito, ativos durante os últimos 6 meses. O arquivo está estruturado na visão cliente.

Fonte: <https://www.kaggle.com/mirichoi0218/insurance>



- (a) Faça a análise exploratória de cada variável individualmente.
- (b) Interprete a média, mediana, Q1, Q2 e Q3 para todas as variáveis.
- (c) Faça o box plot e histograma para todas as variáveis.
- (d) Padronize as variáveis.
- (e) Utilize o método K-médias para 2, 3, 4 e 5 grupos. Qual número de grupos é melhor?
- (f) Caracterize os grupos.
- (g) Crie um plano de ações de marketing para cada grupo.

Arquivo: Marketing_Cartao.xlsx



- Johnson, R. A. e Wichern, D. W. *Applied Multivariate Statistical Analysis*. Prentice-Hall Inc., 6th ed. 2007
- Timm, N.H. *Applied Multivariate Analysis*. Springer-Verlang, 2002



Apêndice – Código R



summary(consumo[, -1]) #Min,
apply(consumo[, -1], 2, sd)

Remove primeira
coluna

Desvio padrão

A função apply () recebe
um data frame ou matriz
como uma entrada e dá a
saída em vetor.
Referência.

Para uma matriz, 1 indica que
queremos analisar as linhas,
e 2 indica colunas.



Função utilizada para
imprimir mais de um gráfico
na mesma tela

Quantidade de colunas

`par(mfrow=c(1,2))`

Quantidade de linhas



Função para padronizar dados

```
#Padronize as variáveis.  
consumo_z <- scale(consumo[, -1])  
head(consumo_z)
```

Remover primeira coluna



Função para cálculo de
distâncias

Remover
primeira coluna

```
distancia <- dist(consumo_z[, -1], method="euclidean")
```

Objeto criado para
armazenar a matriz
de distâncias

Seleção de método do
cálculo da distância



Função que computa o
cluster hierárquico

Método de cálculo da
distância dos grupos

```
clust_single <- hclust(distancia, method="single")  
plot(clust_single, main="Método Single", hang=-1)
```

Título do gráfico

Alinhamento do
dendograma



Função para separar
os grupos

Espessura das
linhas do gráfico

```
rect.hclust(clust_complete, k=3, border=1:5)
```

Quantidade de
grupos



Selecionar somente a coluna
"cluster" do objeto "consumo"

Função para separar os
grupos do cluster hierárquico

```
consumo$cluster <- as.factor(cutree(clust_complete, k=3))  
#Tamanho dos Clusters  
table(consumo$cluster)
```

Função para converter a
variável em categórica

Quantidade de
grupos



```
library(cluster)
```

```
library(factoextra)
```

```
library(gridExtra)
```

Biblioteca que contém os algoritmos de cluster

Biblioteca para visualizar os resultados do cluster k-médias de forma visual

Organiza a saída dos gráficos em forma de matriz (similar ao `par(mfrow())`)



Fixar a semente inicial irá garantir que todas as vezes que o modelo for rodado, sem alteração nos parâmetros, os resultados serão os mesmos



```
set.seed(12345)
```



Função para ajustar o
modelo de cluster pelo
método de k-médias

Critério de parada
utilizado quando o
algoritmo não
finalizar

```
dados.k2 <- kmeans(entregas_z, centers = 2, nstart = 25, iter.max = 100)
```

Quantidade de
grupos

Fixar a
semente inicial



Função para visualizar o cluster de forma gráfica

Título do gráfico

```
#Gráficos  
G1 <- fviz_cluster(dados.k2, geom = "point", data = entregas_z) + ggtitle("k = 2")
```

Formato em que serão exibidas as observações no gráfico



#Criar uma matriz com 4 gráficos

`grid.arrange(G1, G2, G3, G4, nrow = 2)`



Criar uma matriz com todos
os gráficos gerados

