Estatística Aplicada
EAD Ao Vivo

Tema da aula

Análise de Cluster

LABDATA | FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

30/09/2020 - 07/10/2020

ESTATÍSTICA APLICADA EAD AO VIVO

Professora:
Drª Karin Ayumi Tamura

Coordenadores:
Profª Drª Alessandra de Ávila Montini
Profª Dr. Adolpho Walter Pimazoni Canton

Currículo - Prof.ª Drª Karin Ayumi Tamura

FORMAÇÃO ACADÊMICA | EXPERIÊNCIA PROFISSIONAL

Profª Dra.
Karin Ayumi Tamura

Contato: karin.tamura@fia.com.br

- **FORMAÇÃO ACADÊMICA:** Pós-doutora (2015), Doutora (2012), mestre (2007) e bacharel (2003) em Estatística pelo Instituto de Matemática e Estatística da USP, tendo como área de pesquisa modelos de regressão, análise multivariada de dados e algoritmos de *machine learning*.
- **ATUAÇÃO PROFISSINAL:** Foi *Head* de *Analytics* por 14 anos, e atualmente é Conselheira Executiva e *Head* de Inovação na *Marketdata Solutions*, uma empresa do grupo WPP, e Professora Doutora no LABDATA FIA.
- **HISTÓRICO:** Atuação no mercado por 17 anos, com experiência profissional no segmento bancário (Bradesco) e consultoria *(Marketdata Solutions)*. Atuou como docente em cursos de pós-graduação (2010-16) no LABDATA FIA e ABEMD. Especialista em Estatística e *Advanced Analytics* trabalhando em projetos de diversos segmentos do mercado. Participante de congressos nacionais e internacionais voltados a área de Estatística, Dados e Algoritmos de *Machine Learning*.

"Tenho duas paixões no meu trabalho: dados e pessoas. Meu objetivo como professora é integrar a visão do mercado com as técnicas e tecnologias de análise de dados, por meio de uma atuação humanista no ensino aos alunos"

Projetos atendidos

FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

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

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

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**.


Profª Drª Alessandra Montini

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 Programático do Curso

21 AULAS AO VIVO COM PROFA. KARIN | 27 PLANTÕES AO VIVO COM PROF. STEPHAN, 7 LISTAS DE EXERCÍCIOS E EAD VIDEO AULA EM PYTHON

Dia	Mês	Aula	EAD Ao Vivo	Plantão Prof. Stephan
5	Agosto	Introdução ao Curso e Análise Exploratória de Dados	Aula Prof. Karin	06/ago
12	Agosto	Análise Exploratória de Dados	Aula Prof. Karin	13/ago
19	Agosto	Análise Exploratória de Dados - Introdução ao R	Aula Prof. Karin	20/ago
26	Agosto	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	27/ago
2	Setembro	Regressão Linear Simples	Aula Prof. Karin	03/set
9	Setembro	Regressão Linear Simples e Múltipla	Aula Prof. Karin	10/set
16	Setembro	Regressão Linear Simples e Múltipla	Aula Prof. Karin	17/set
23	Setembro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	24/set
30	Setembro	Análise de Cluster	Aula Prof. Karin	01/out
7	Outubro	Análise de Cluster	Aula Prof. Karin	08/out
14	Outubro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	15/out
21	Outubro	Arvore de Decisão	Aula Prof. Karin	22/out
28	Outubro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	29/out
4	Novembro	Regressão Logística	Aula Prof. Karin	05/nov
11	Novembro	Regressão Logística	Aula Prof. Karin	11/nov
18	Novembro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	19/nov
25	Novembro	estudo de caso	Aula Prof. Karin	26/nov
2	Novembro	estudo de caso	Aula Prof. Karin	30/dez
9	Dezembro	estudo de caso	Aula Prof. Karin	10/dez
16	Dezembro	Análise de Série Temporal - modelo auto regressivo	Aula Prof. Karin	17/dez
23	Dezembro	Lista de Exercícios em Sala de Aula (Frequência Liberada - véspera Natal)	-	-
Recesso Escolar		EAD - INTRODUÇÃO AO PYTHON EAD - INTRODUÇÃO AO PYTHON	EAD Video Aula (8 horas)	- -
6	Janeiro	Modelos estatísticos em Python	Aula Prof. Karin	07/jan
13	Janeiro	Modelos estatísticos em Python	Aula Prof. Karin	14/jan
20	Janeiro	Modelos estatísticos em Python	Aula Prof. Karin	20/jan
27	Janeiro	Introdução a Big Data - Aplicações de Machine Learning e Deep Learning	Aula Prof. Karin	28/jan
3	Fevereiro	Aplicações de Machine Learning	Aula Prof. Karin	04/fev
10	Fevereiro	Aplicações de Machine Learning	Aula Prof. Karin	11/fev
17	Fevereiro	Lista de Exercícios (Frequência Liberada - quarta de cinzas)	-	18/fev
24	Fevereiro	EXERCICIOS DE REVISÃO - EAD (19hs e 23hs - com presença obrigatória)	-	24/fev
3	Março	Prova (Plataforma On Line: 19hs e 23hs)	-	

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 para casa
   i. FIXAÇÃO: Método Hierárquico e Dendrograma
   ii. CASE: Hábitos Alimentares

1. Introdução

Case Encarteiramento de clientes

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

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

Aplicação
SAC e Ouvidoria.

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

Aplicação
Gestão de Pessoas.

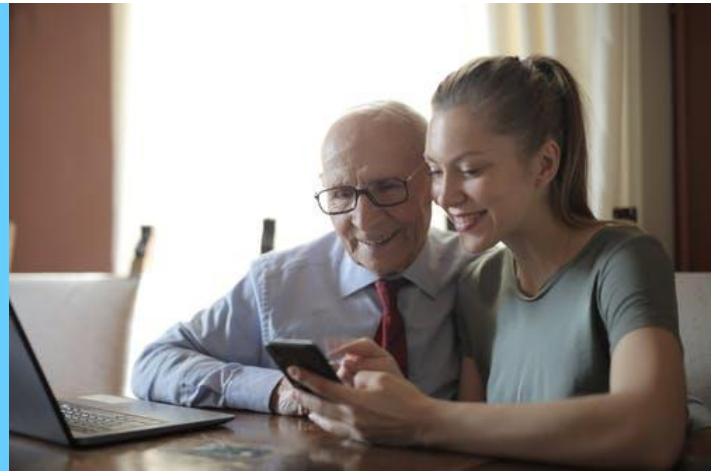
Case Hábitos Alimentares

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

Aplicação
Áreas de Saúde & Nutrição.

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

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

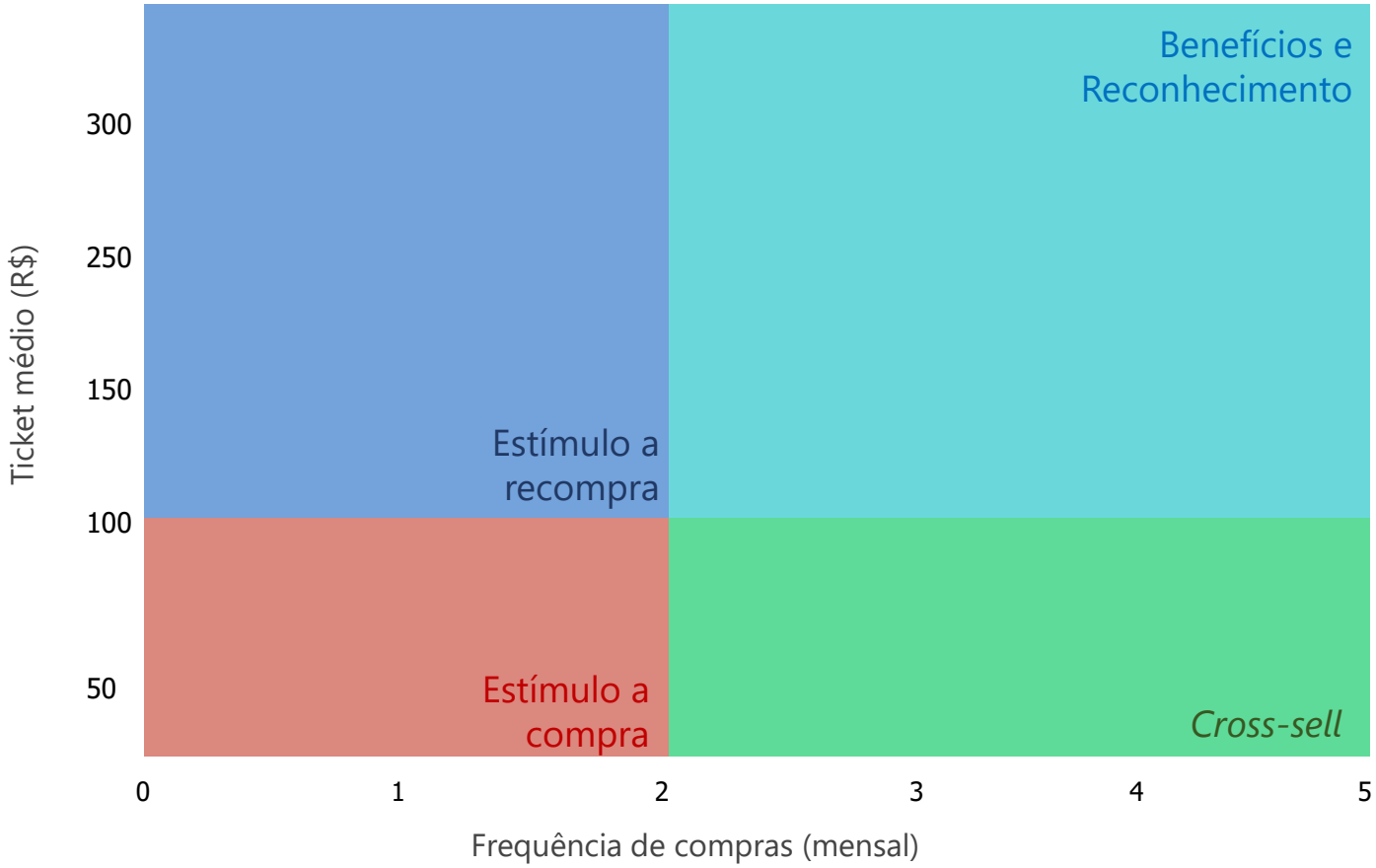
Aplicação
Marketing & CRM.

Exemplo
Estratégia de reconhecimento e relacionamento com clientes de acordo com sua transacionalidade (Recência, Frequência e Valor).

Aplicação
Marketing & CRM.

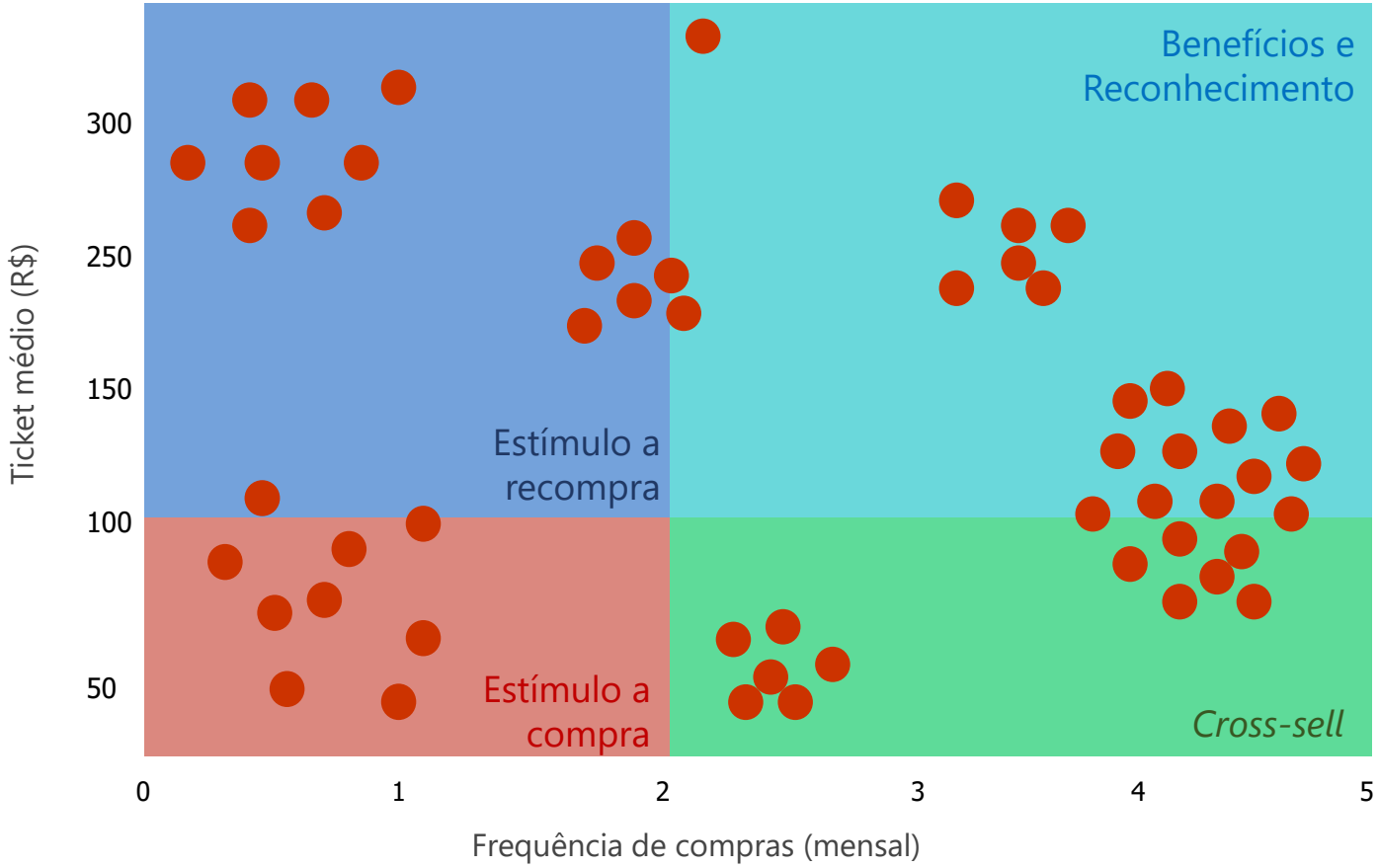
Case: Varejo (Frequência e Valor)

Estratégias de reconhecimento e relacionamento segmentadas para **4 grupos** de Transacionalidade (Frequência e Valor).



Case: Varejo (Frequência e Valor)

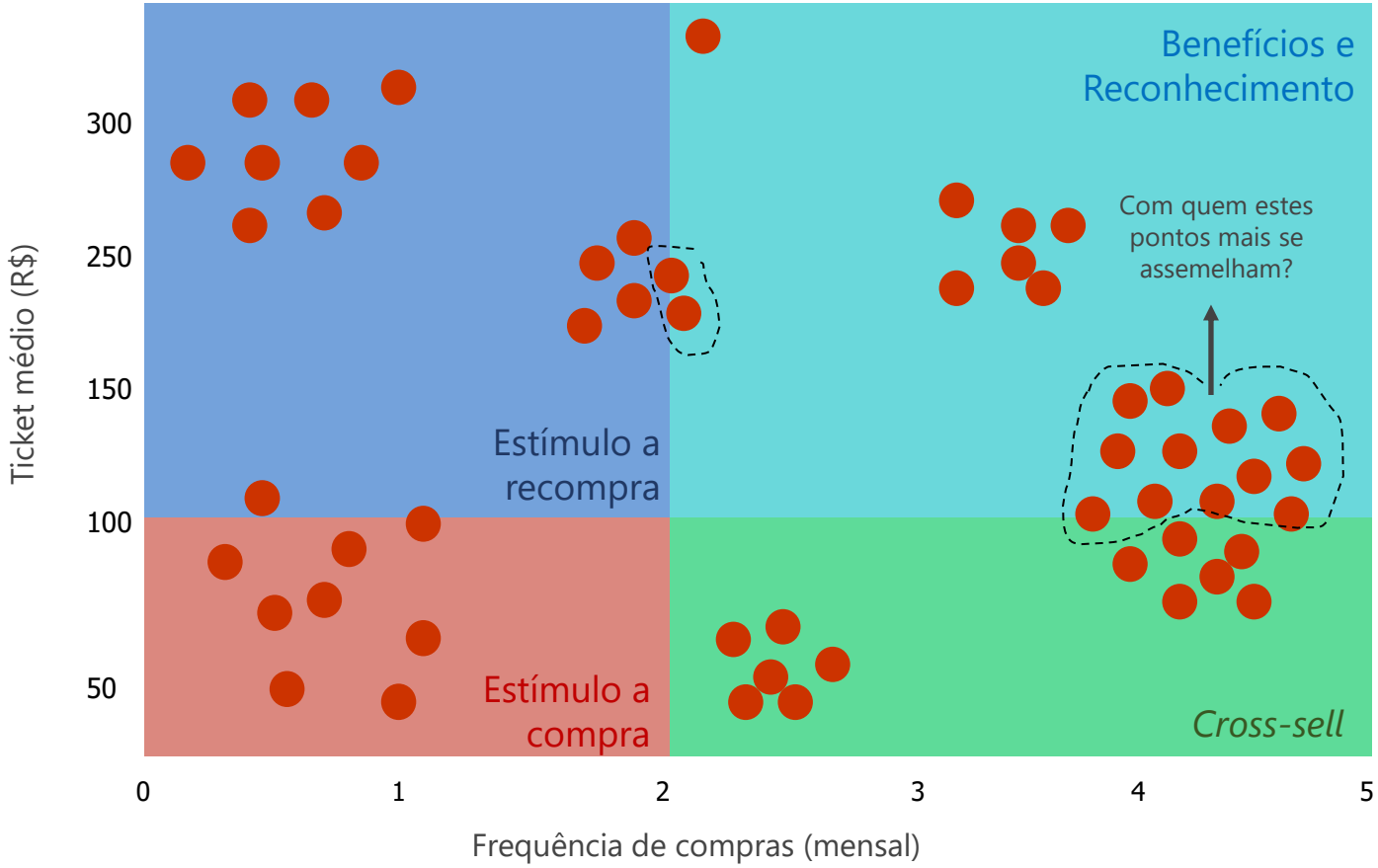
Estratégias de reconhecimento e relacionamento segmentadas para **4 grupos** de Transacionalidade (Frequência e Valor).



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

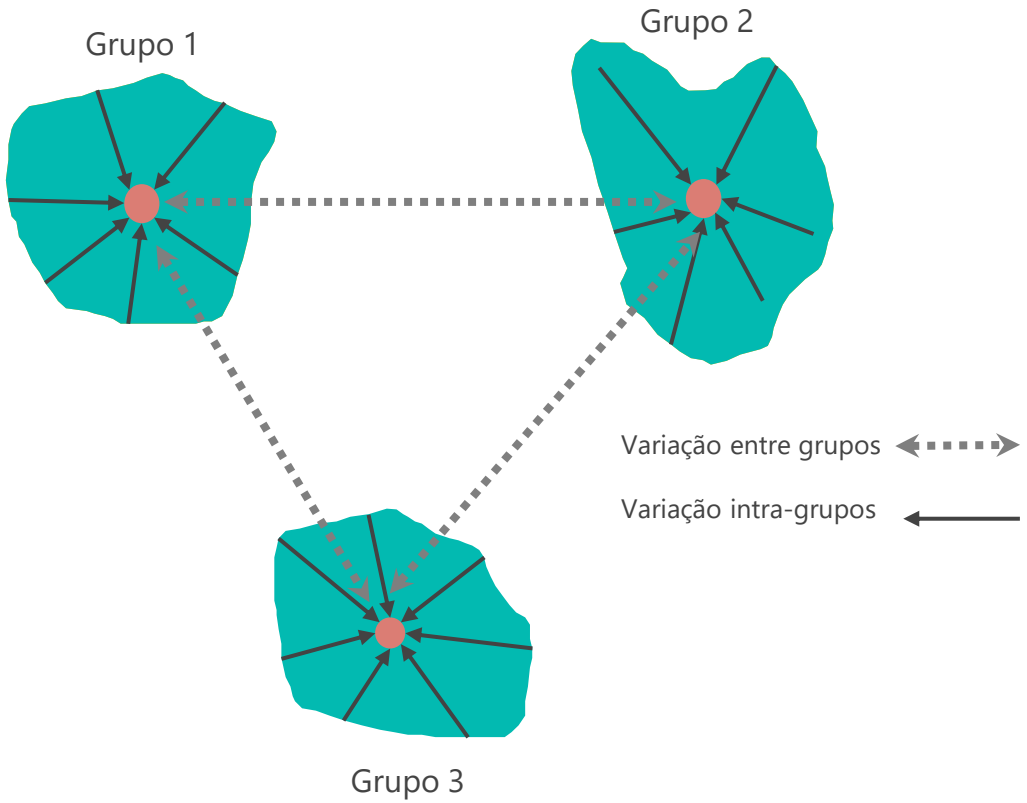
Case: Varejo (Frequência e Valor)

Estratégias de reconhecimento e relacionamento segmentadas para **4 grupos** de Transacionalidade (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*



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 **heterogêneo** entre os grupos.

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

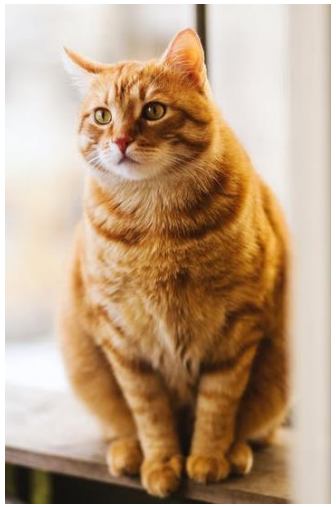
Segmentação

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?

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

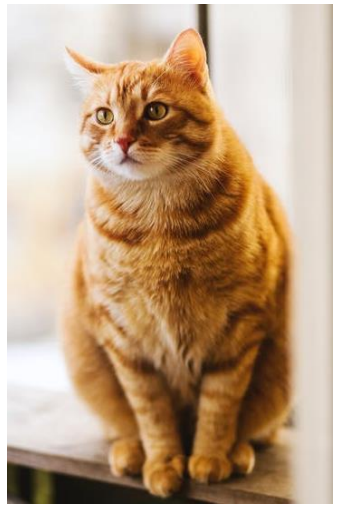


Como identificamos indivíduos (observações) semelhantes?

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.

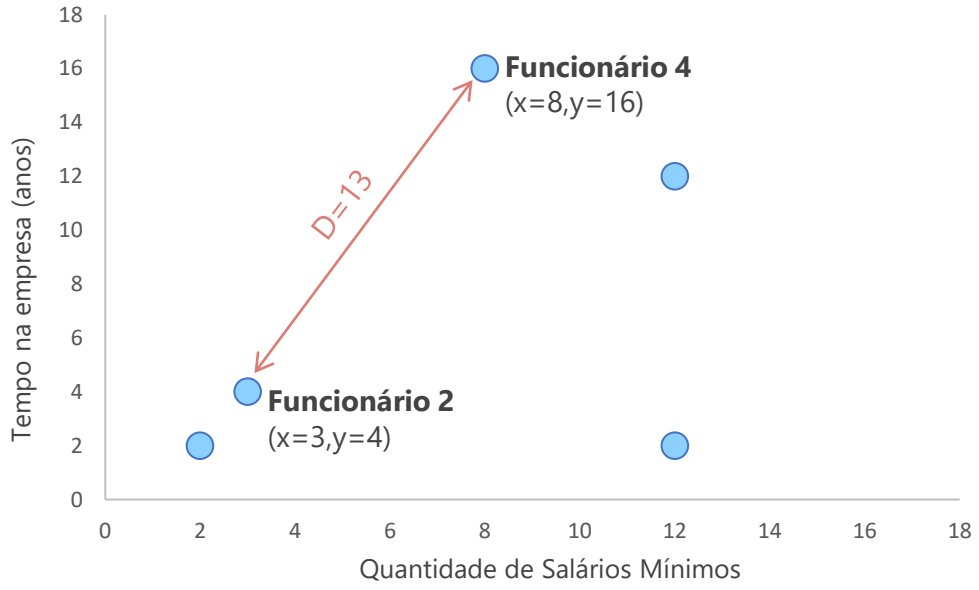
Segmentação

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

Distância Euclidiana

Na Análise de *Cluster* as observações são agrupadas de acordo com medidas de similaridade.

Um **critério de dissimilaridade** (quanto menor o valor mais parecido) que pode ser considerado para agrupar observações é a **Distância Euclidiana**.



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

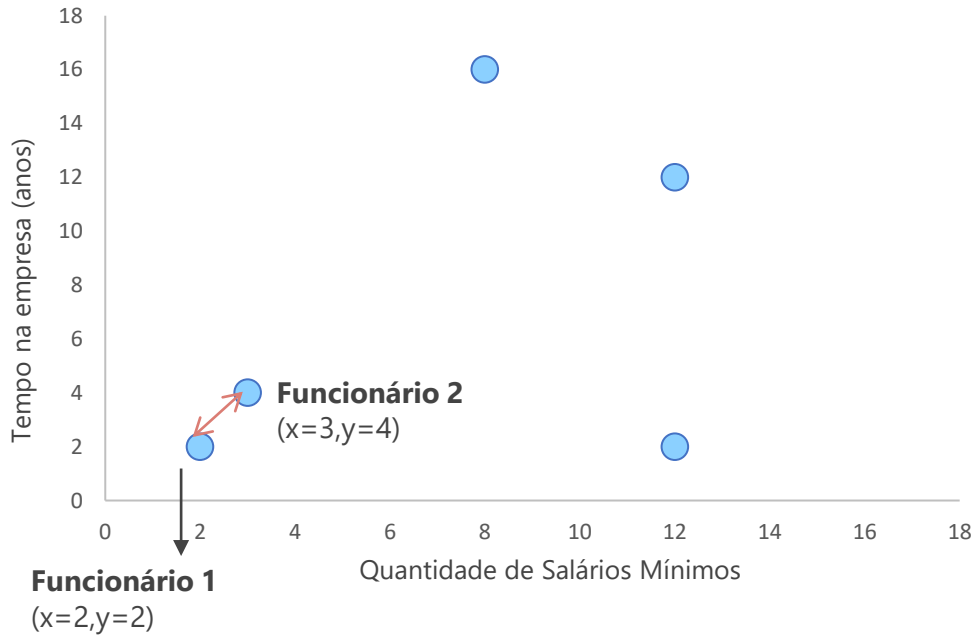
$$D^2 = (8-3)^2 + (16-4)^2 = 5^2 + 12^2 = 169$$

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

Exercício: Calcule a Distância Euclidiana

Calcule a distância Euclidiana entre os funcionários 1 e 2.

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



A **Distância Euclidiana (D)** entre os funcionários 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 funcionários estão, então o funcionário 2 está mais próximo do funcionário 1 e mais distante do funcionário 4.

Matriz de distâncias

Fazendo o cálculo da distância (euclidiana) entre todas as observações, obtém-se uma **matriz de distâncias**, que é **simétrica.**

Matriz de distâncias

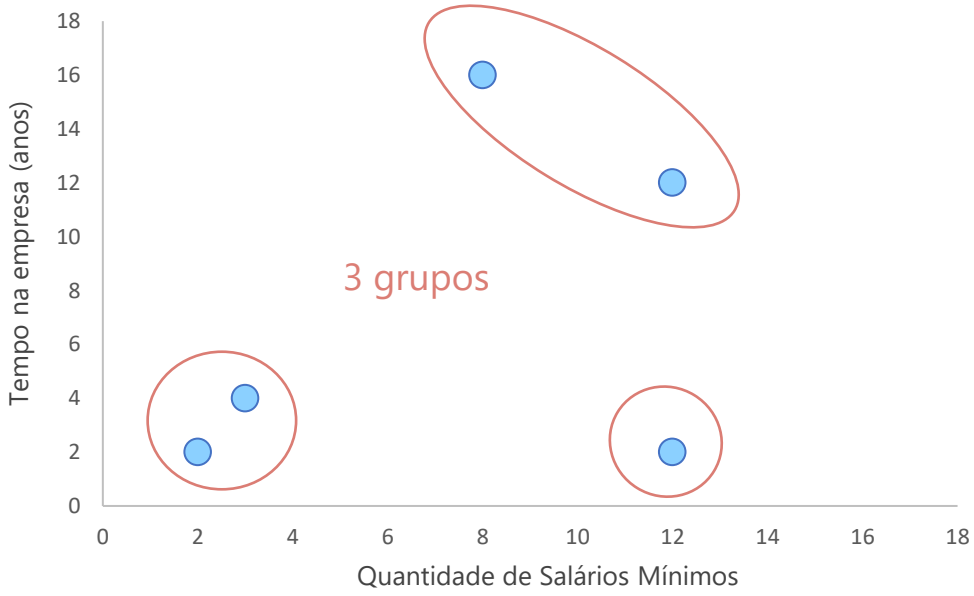
	1	2	3	4	5
1		2,24	14,14	15,23	10,00
2			12,04	13,00	9,22
3				5,66	10,00
4					14,56
5					

Matriz de distâncias

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 funcionários.

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

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

Método *Single* (vizinho mais próximo)

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)

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					

Menor distância

Passo 1: juntar 1+2

	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)

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					

Menor distância

Passo 1: juntar 1+2

	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, agora vamos agrupar as informações dos funcionários 1 e 2, por meio do **MÍNIMO**:

Distância entre 1 e 3 = 14,14	Distância entre 1 e 4 = 15,23	Distância entre 1 e 5 = 10,00
Distância entre 2 e 3 = 12,04	Distância entre 2 e 4 = 13,00	Distância entre 2 e 5 = 9,22
MÍNIMO é 12,04	**MÍNIMO** é 13,00	**MÍNIMO** é 9,22

Método *Single* (vizinho mais próximo)

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					

Menor distância &
agrega pelo MÍNIMO

Passo 1: juntar 1+2

	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					

	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)

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					

Menor distância & agrega pelo MÍNIMO

Passo 1: juntar 1+2

	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					

Menor distância

	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)

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					

Menor distância & agrega pelo MÍNIMO

Passo 1: juntar 1+2

	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					

Menor distância

Passo 2: juntar 3+4

	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

Agregamos pelo MÍNIMO

Dado que a menor distância é 5,66, agora vamos agrupar as informações dos funcionários 3 e 4, por meio do **MÍNIMO**:

Distância entre 1+2 e 3 = 12,04	Distância entre 3 e 5 = 10,00
Distância entre 1+2 e 4 = 13,00	Distância entre 4 e 5 = 14,56
MÍNIMO é 12,04	**MÍNIMO** é 10,00

Método *Single* (vizinho mais próximo)

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					

Menor distância & agrega pelo MÍNIMO

Passo 1: juntar 1+2

	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					

Menor distância & agrega pelo MÍNIMO

Passo 2: juntar 3+4

	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'+5

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

Método *Single* (vizinho mais próximo)

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					

Menor distância & agrega pelo MÍNIMO

Passo 1: juntar 1+2

	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					

Menor distância & agrega pelo MÍNIMO

Passo 2: juntar 3+4

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

Menor distância

Passo 3: juntar '1+2'+5

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

Método *Single* (vizinho mais próximo)

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					

Menor distância & agrega pelo MÍNIMO

Passo 1: juntar 1+2

	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					

Menor distância & agrega pelo MÍNIMO

Passo 2: juntar 3+4

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

Menor distância & agrega pelo MÍNIMO

Passo 3: juntar '1+2'+5

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

Dado que a menor distância é 9,22, agora vamos agrupar as informações dos funcionários 1+2 e 5, por meio do **MÍNIMO**:

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)

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					

Menor distância & agrega pelo MÍNIMO

Passo 1: juntar 1+2

	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					

Menor distância & agrega pelo MÍNIMO

Passo 2: juntar 3+4

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

Menor distância & agrega pelo MÍNIMO

Passo 3: juntar '1+2'+5

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

Menor distância & agrega pelo MÍNIMO

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

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

Método *Single* (vizinho mais próximo)

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					

Menor distância & agrega pelo MÍNIMO

Passo 1: juntar 1+2

	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					

Menor distância & agrega pelo MÍNIMO

Passo 2: juntar 3+4

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

Menor distância & agrega pelo MÍNIMO

Passo 3: juntar '1+2'+5

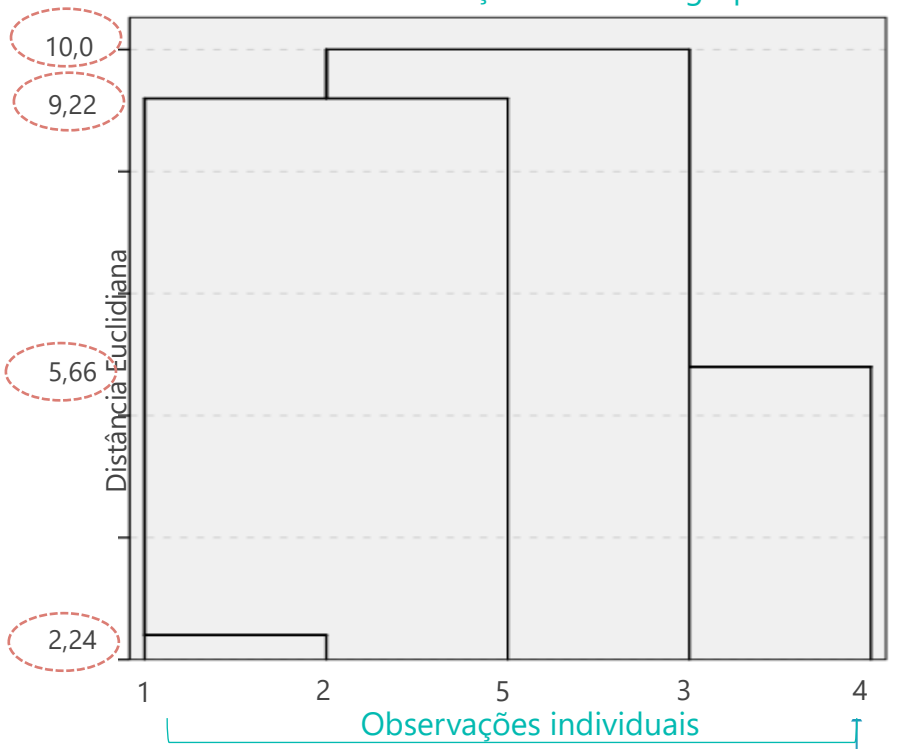
	1+2	3+4	5
1 + 2		12,04	9,22
3+4			10,00
5			

Menor distância & agrega pelo MÍNIMO

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

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

Todas as observações em único grupo

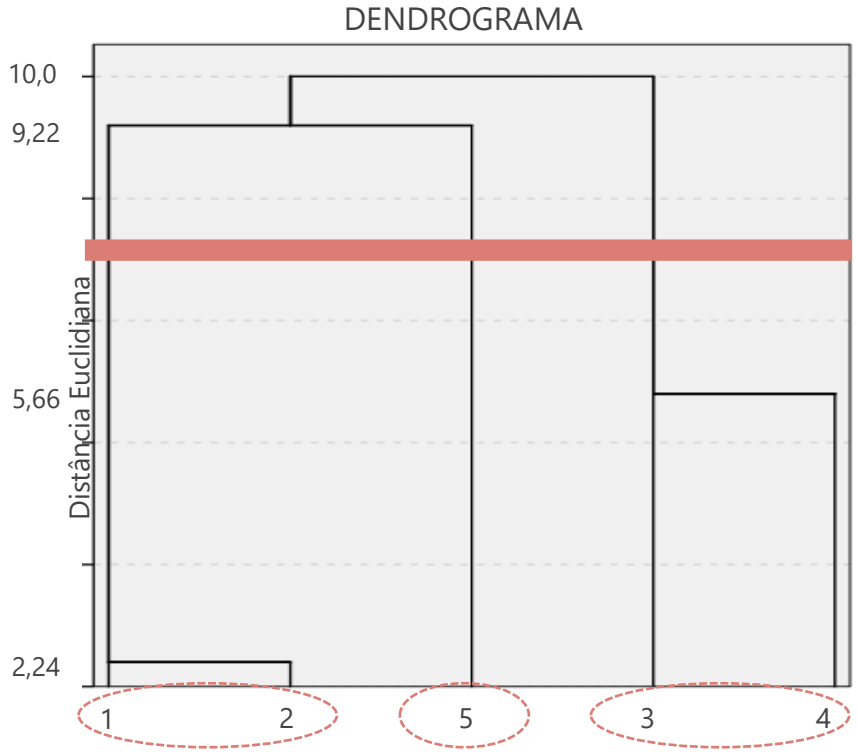


Dendrograma

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

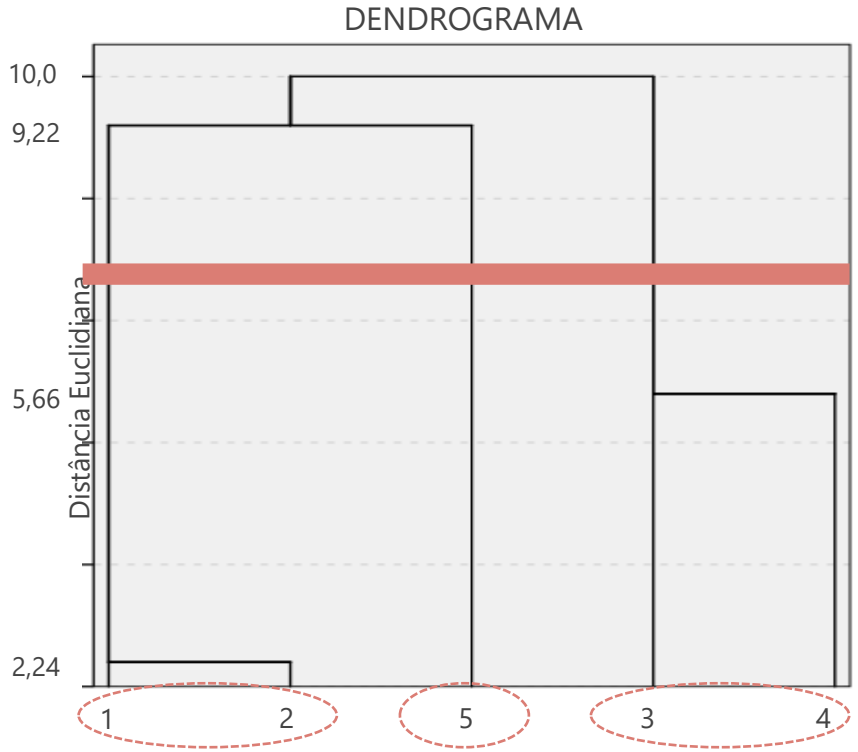
Como a distância entre 9,22 e 5,66 é grande, pode-se sugerir separar os grupos em uma distância superior a 5,657 e inferior a 9,220.

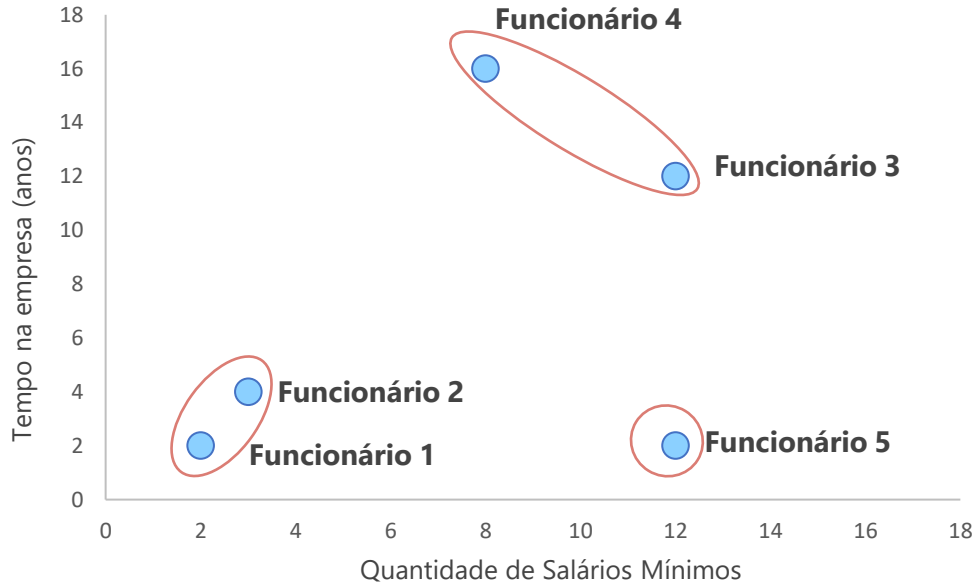
A linha vermelha representa a separação, e abaixo dela a quantidade de grupos formados, no exemplo, 3 grupos.

Dendrograma

O dendrograma (à esquerda) sugere 3 grupos, assim como visto 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 para visualizar a formação dos grupos.





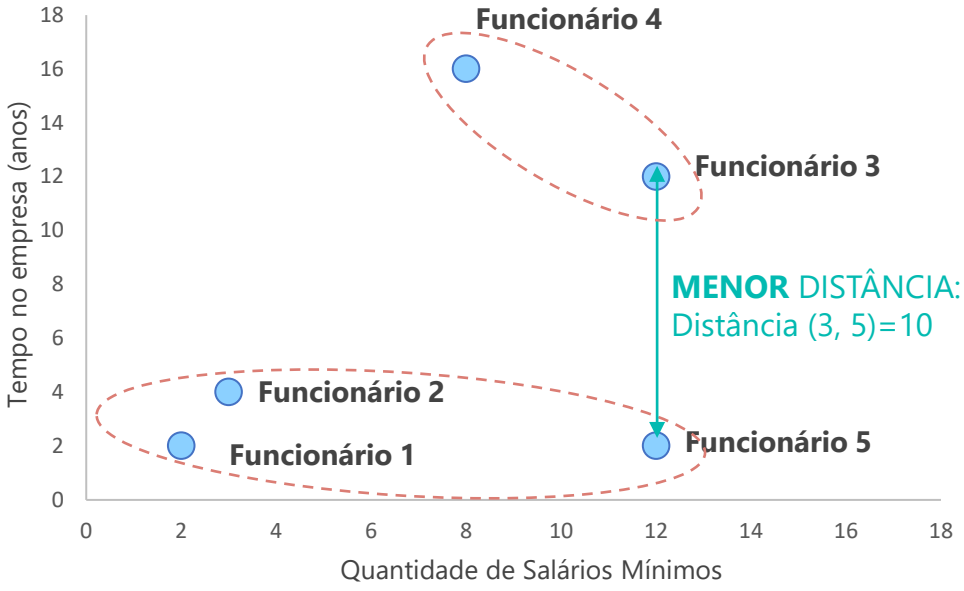
Discussão entre os métodos

Dado a escolha de uma medida de dissimilaridade (p.e. a Distância Euclidiana), precisamos escolher um **critério para agregar as observações**, como as apresentadas a seguir:

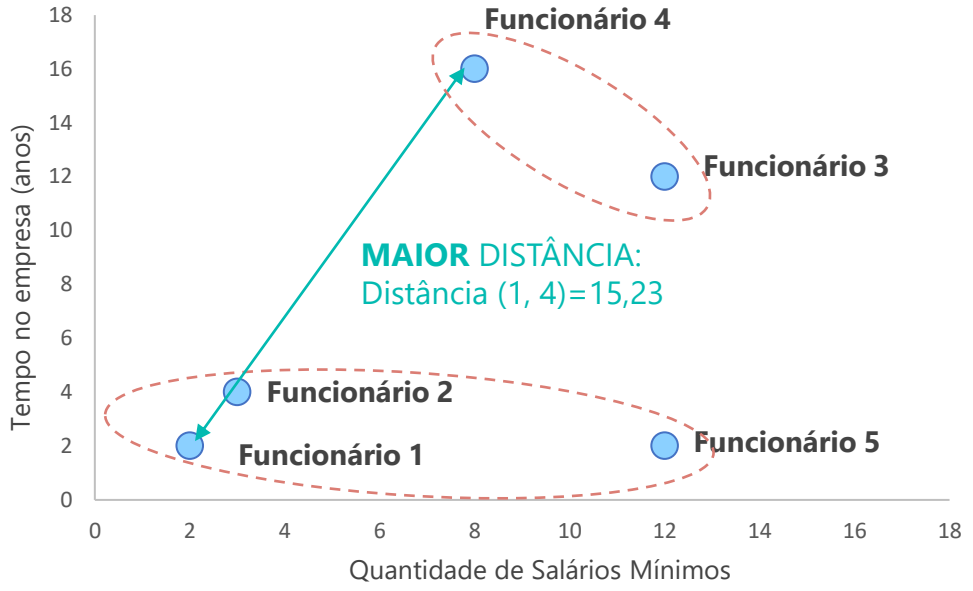
- ***Single* (vizinho mais próximo):** define-se como o MÍNIMO da distância entre um elemento do outro.
- ***Complete* (vizinho mais longe):** define-se como o MÁXIMO da distância entre um elemento do outro.

Método *Single e Complete*

Single (critério do MÍNIMO)



Complete (critério do MÁXIMO)



Nota: Mais detalhes em Johnson (2007), página 680 - 695.

Método *Complete* (vizinho mais longe)

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					

Menor distância

Passo 1: juntar 1+2

	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, agora vamos agrupar as informações dos candidatos 1 e 2, por meio do **MÁXIMO**:

Distância entre 1 e 3 = 14,14	Distância entre 1 e 4 = 15,23	Distância entre 1 e 5 = 10,00
Distância entre 2 e 3 = 12,04	Distância entre 2 e 4 = 13,00	Distância entre 2 e 5 = 9,22
MÁXIMO é 14,14	**MÁXIMO** é 15,23	**MÁXIMO** é 10,00

Case: Hábitos Alimentares

Os dados são de uma pesquisa de consumo de alimentos em 25 países da Europa. Nove grupos de comida foram analisados: carne vermelha, carne branca, ovos, leite, peixes, cereais, carboidratos, grãos, frutas e vegetais. Os dados foram obtidos de DASL *(The Data and Story Library*).
O objetivo do estudo é agrupar os países segundo o comportamento de hábitos alimentares semelhantes, e investigar os hábitos alimentares com indicadores de longevidade e doenças crônicas de cada grupo de países.

Vamos fazer juntos?

RStudio

(a) Abra o banco de dados *Consumo_Alimentos.txt* no R.
(b) Faça uma análise exploratória da base de dados. Comente sobre a variabilidade dos dados.
(c) Calcule a matriz de distâncias euclidianas entre os 25 países.
(d) Faça a análise de agrupamento usando os 2 métodos apresentados, escolha um dos métodos e justifique a quantidade de grupos após a análise do Dendrograma.
(e) Analise as características de cada grupo pela análise do *Box Plot*. Comente os resultados.

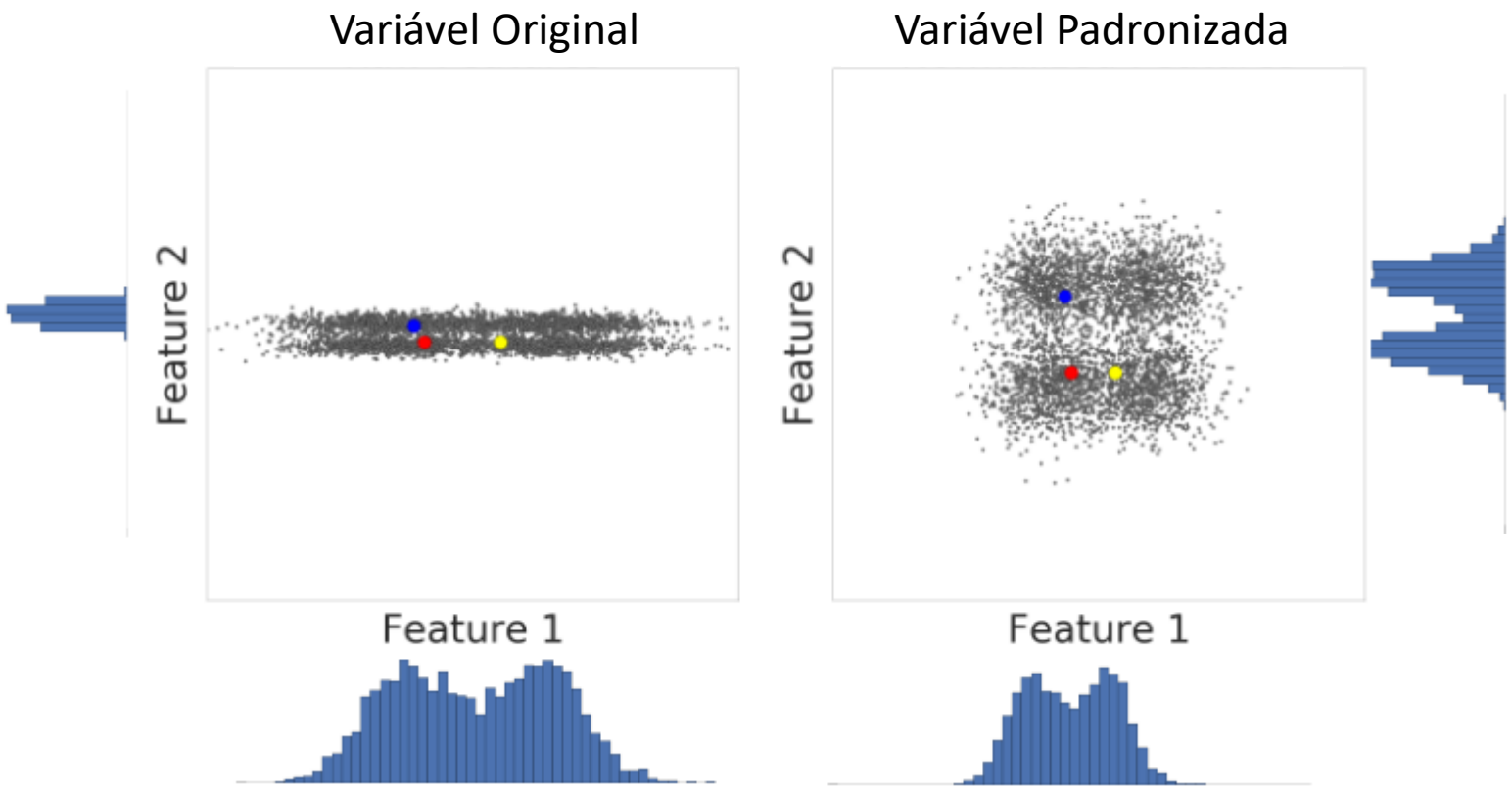
Discussão entre os métodos

- O método *complete* tende a formar grupos mais homogêneos do que o método *single*, pois uma distância pequena entre os dois grupos implica na proximidade de todos os elementos desses grupos.
- O método *single* tende a formar grupos mais heterogêneos, pois apesar da distância ser pequena, há elementos que diferem muito entre si.

Observação: Os métodos hierárquicos podem fornecer agrupamentos diferentes. Por isso, é importante testar alguns deles e verificar qual o método que forneceu o agrupamento que está mais de acordo com a aplicação do problema.

3. Padronização das variáveis

Z-score



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

Z-score

- Variáveis com maior dispersão (maior desvio padrão) tem 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 igual desvio padrão para todas as varáveis.
- Para se obter uma variável padronizada deve-se subtrair de cada valor a média e dividir pelo desvio padrão.

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

Case: Hábitos Alimentares

Os dados são de uma pesquisa de consumo de alimentos em 25 países da Europa. Nove grupos de comida foram analisados: carne vermelha, carne branca, ovos, leite, peixes, cereais, carboidratos, grãos, frutas e vegetais. Os dados foram obtidos de DASL (*The Data and Story Library*).
O objetivo do estudo é agrupar os países segundo comportamento de hábitos alimentares semelhantes, e investigar os hábitos alimentares com indicadores de longevidade e doenças crônicas de cada grupo de países.

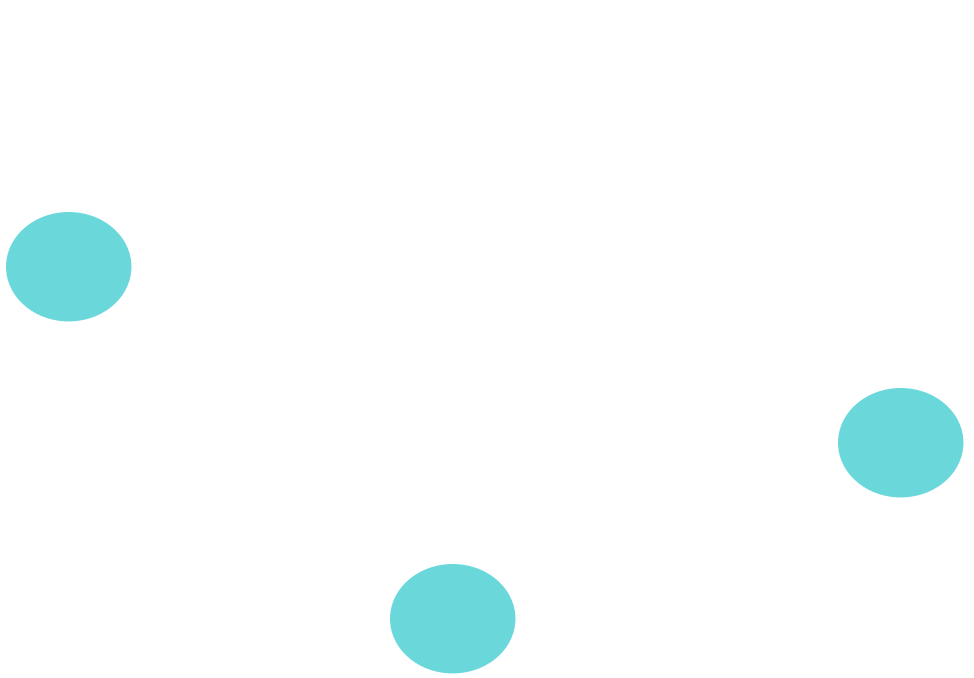
Vamos fazer juntos?

RStudio

(a) Abra o banco de dados *Consumo_Alimentos.txt* no R.
(b) Faça uma análise exploratória da base de dados. Comente sobre a variabilidade dos dados.
(c) Calcule a matriz de distâncias euclidianas entre os 25 países.
(d) Faça a análise de agrupamento usando os 2 métodos apresentados, escolha um dos métodos e justifique a quantidade de grupos após a análise do Dendrograma.
(e) Analise as características de cada grupo pela análise do *Box Plot*. Comente os resultados.
(f) Padronize as variáveis e refaça os itens (d) e (e).

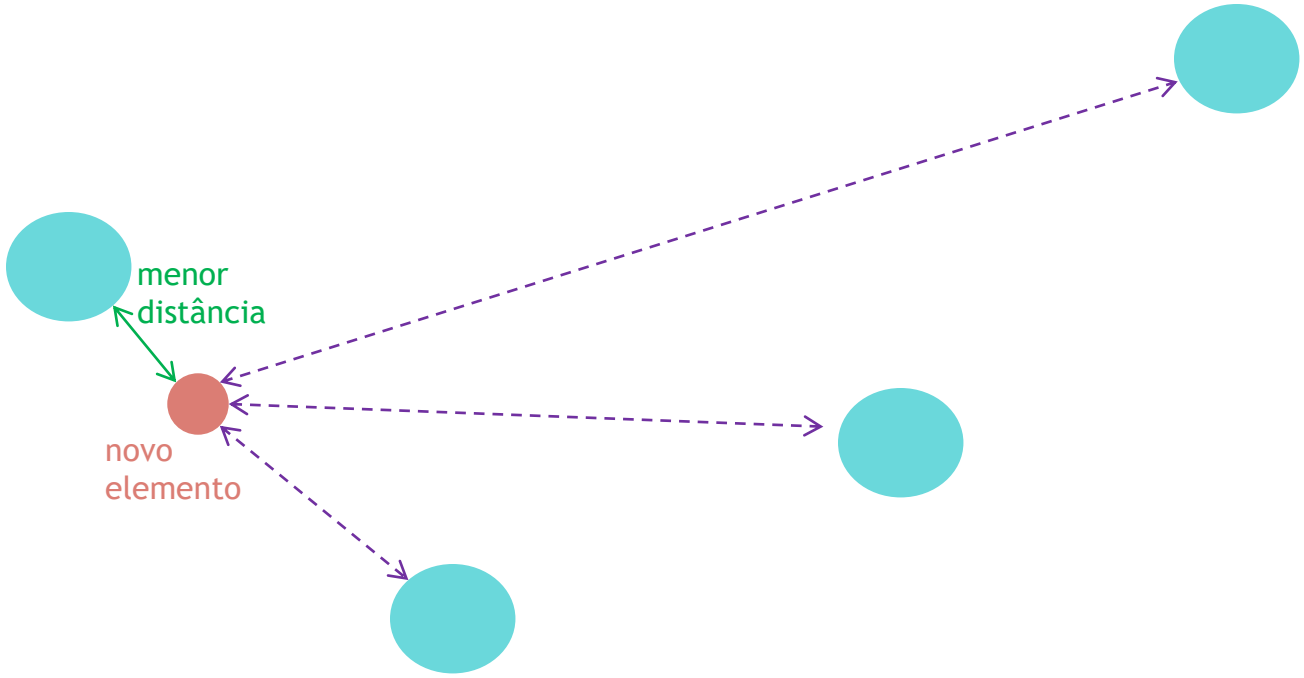
4. Método de Partição: K-médias

Processo Iterativo



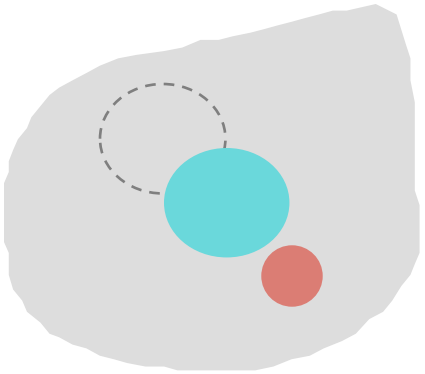
O algoritmo começa com a formação de uma partição inicial de k centroides (sementes) aleatórias. Neste caso, k=4.

Processo Iterativo



Dado um novo elemento, calcula-se a distância (euclidiana) mais próxima deste elemento para as 4 centroides. O novo elemento é agrupado com a centroide mais próxima.

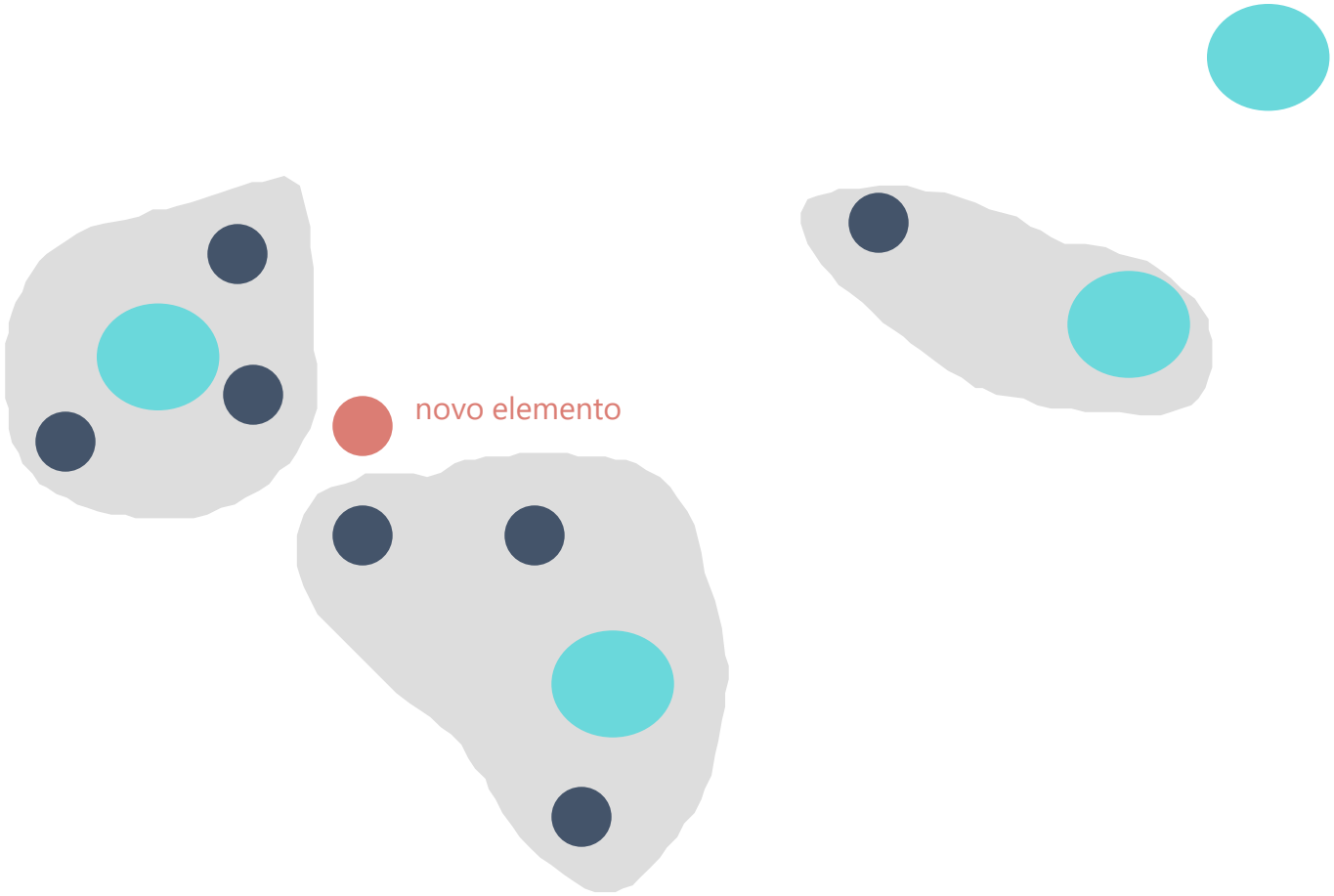
Processo Iterativo



nova centroide

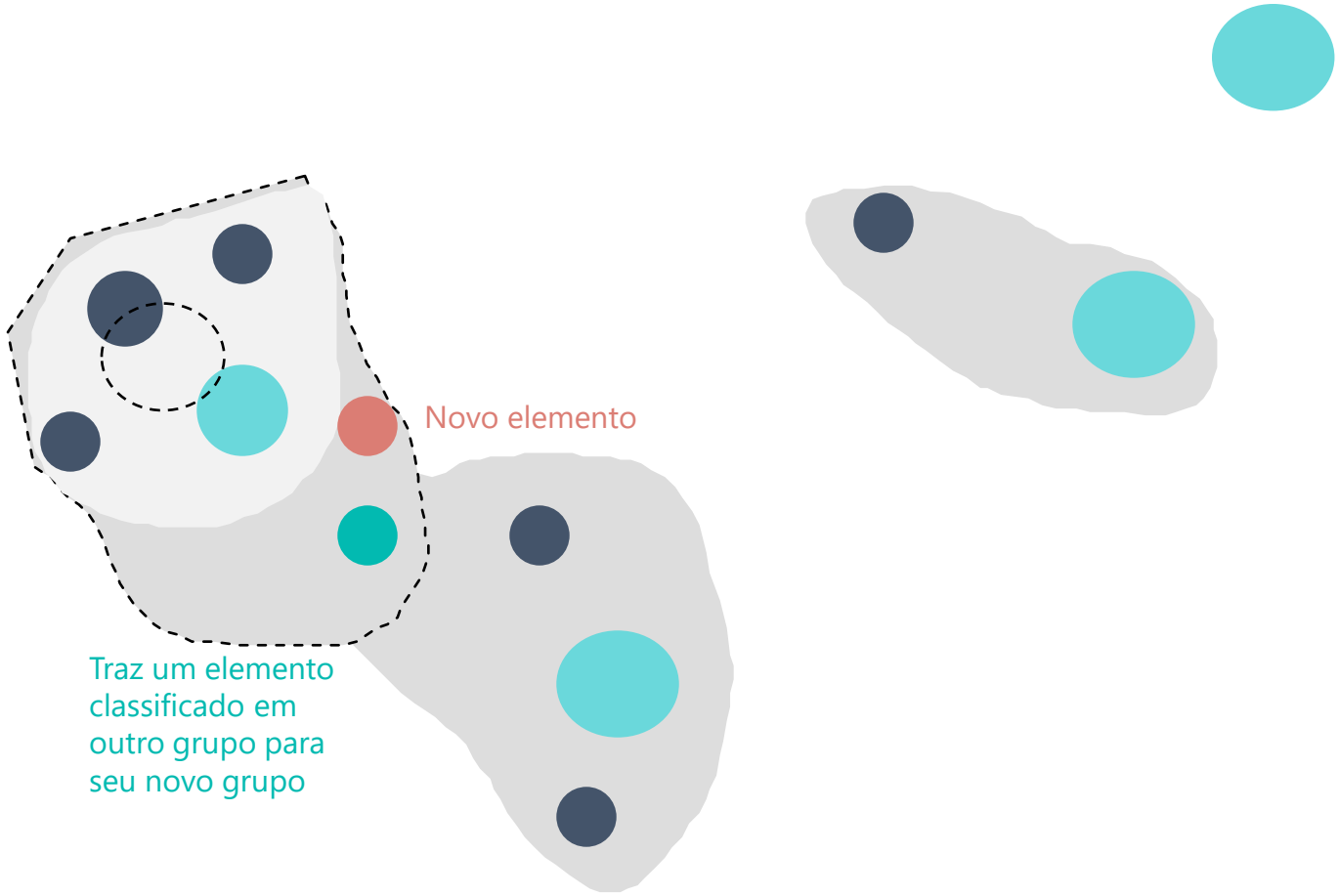
A nova centroide é recalculada.

Processo Iterativo



A cada passo, as observações são agrupadas no *cluster* com a centroide mais próxima, com subsequentes recálculos das centroides.

Processo Iterativo



As observações são agrupadas nos *clusters* até que as partições encontradas maximizem critério de homogeneidade dentro do grupo.

K-médias

Utiliza-se de um procedimento de aproximação e por isso pode ser usado em grandes bancos de dados.

O número de *clusters* (k) precisa ser previamente definido.

Método de Partição: ***K-means*** **(K-médias)**

- A centroide de um grupo é definida como a média das distâncias de seus elementos.
- É um processo iterativo no qual, a cada passo, os elementos são agrupados no *cluster* com a centroide mais próxima, com subsequentes recálculos dos centros.
- As observações são agrupadas nas centroides até que as partições encontradas satisfaçam o critério de qualidade adotado.

Processo Iterativo

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

Caso as sementes iniciais não estejam adequadas, realiza-se um processo iterativo.

Case: Hábitos Alimentares

Os dados são de uma pesquisa de consumo de alimentos em 25 países da Europa. Nove grupos de comida foram analisados: carne vermelha, carne branca, ovos, leite, peixes, cereais, carboidratos, grãos, frutas e vegetais. Os dados foram obtidos de DASL (*The Data and Story Library*).
O objetivo do estudo é agrupar os países segundo comportamento de hábitos alimentares semelhantes, e investigar os hábitos alimentares com indicadores de longevidade e doenças crônicas de cada grupo de países.

Vamos fazer juntos?

RStudio

(a) Abra o banco de dados *Consumo_Alimentos.txt* no R.
(b) Faça uma análise exploratória da base de dados. Comente sobre a variabilidade dos dados.
(c) Calcule a matriz de distâncias euclidianas entre os 25 países.
(d) Faça a análise de agrupamento usando os 2 métodos apresentados, escolha um dos métodos e justifique a quantidade de grupos após a análise do Dendrograma.
(e) Analise as características de cada grupo pela análise do *Box Plot*. Comente os resultados.
(f) Padronize as variáveis e refaça os itens (d) e (e).
(g) Rode o método K-médias, utilizando o k encontrado no item (f).

5. Exercícios para casa

5. Exercícios para casa

DATA DE ENTREGA 18/10/2020 | 1 EXERCÍCIO-CASE e 1 EXERCÍCIO DE FIXAÇÃO

i. FIXAÇÃO: Método Hierárquico (3,0 ponto)
ii. CASE: Hábitos Alimentares (7,0 pontos)

Instruções importantes:

- A lista vale nota (0-10) e deve ser entregue até 18/10/2020. Lista entregue até 25/10/2020 valerá 80% da nota. Posteriormente, não será mais aceita a lista para correção. Não serão aceitas listas parciais.
- O exercício será considerado como "realizado", quando tiver, além das análises, a interpretação do resultados.
- Soluções técnicas "elegantes e mais completas" serão considerados como ponto extra para o aluno (+1,0 na lista geral).
- Caso o aluno tire nota >10, considerando os pontos extras, os pontos extras poderão ser acumulados para listas seguintes, sendo a média geral de todas as listas realizadas no curso, com valor máximo igual a 10.

BOM ESTUDO ☺

5.i. Método Hierárquico e Dendrograma

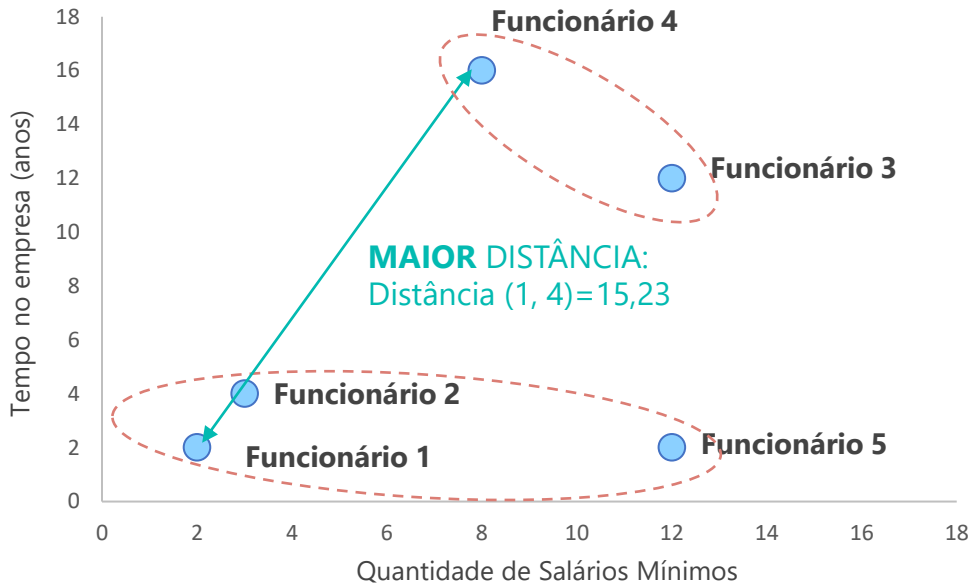
EXERCÍCIO DE FIXAÇÃO | NÃO É NECESSÁRIO UTILIZAR O R PARA RESOLVER O PROBLEMA

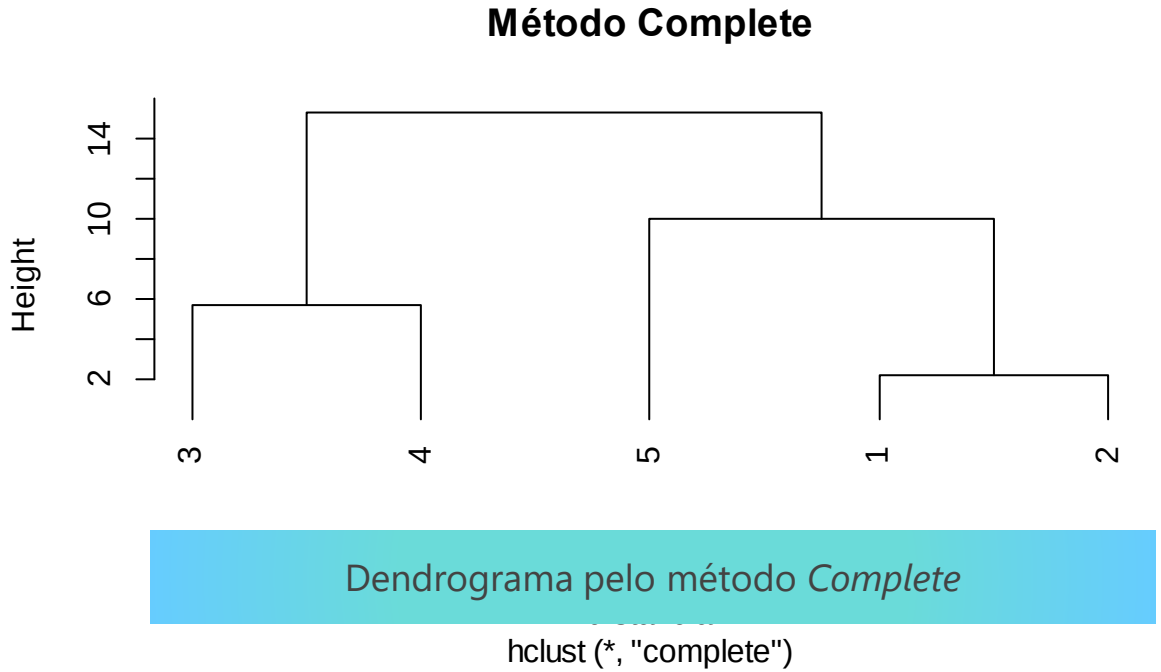
Reproduza os passos descritos nos slides 31 a 41, agora para método *Complete*.
Utilize a mesma matriz de distâncias euclidianas apresentada no slide 33, e obtenha passo-a-passo os valores das distâncias máximas a cada passo de agrupamento, conforme dendrograma abaixo.

Passo 0: matriz de distâncias (Slide 31)

	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					

Exemplo: *Complete* (critério do MÁXIMO) – Slide 45





Dendrograma pelo método *Complete*

5.ii. Case: Hábitos Alimentares

Os dados são de uma pesquisa de consumo de alimentos em 25 países da Europa. Nove grupos de comida foram analisados: carne vermelha, carne branca, ovos, leite, peixes, cereais, carboidratos, grãos, frutas e vegetais. Os dados foram obtidos de DASL (*The Data and Story Library*).
O objetivo do estudo é agrupar os países segundo comportamento de hábitos alimentares semelhantes, e investigar os hábitos alimentares com indicadores de longevidade e doenças crônicas de cada grupo de países. **Os itens (a)-(g) já foram realizados em sala, agora precisam apenas ser complementados com seus respectivos comentários e conclusões.**

(a) Abra o banco de dados *Consumo_Alimentos.txt* no R.
(b) Faça uma análise exploratória da base de dados. Comente sobre a variabilidade dos dados.
(c) Calcule a matriz de distâncias euclidianas entre os 25 países.
(d) Faça a análise de agrupamento usando os 2 métodos apresentados, escolha um dos métodos e justifique a quantidade de grupos após a análise do Dendrograma.
(e) Analise as características de cada grupo pela análise do *Box Plot*. Comente os resultados.
(f) Padronize as variáveis e refaça os itens (d) e (e).
(g) Rode o método K-médias, utilizando o k encontrado no item (f).
(h) Realize a comparação dos grupos encontrados pelos métodos
    i. *Complete* sem padronização
    ii. *Complete* com padronização
    iii. K-médias com padronização

    Qual agrupamento você escolheria? Descreva o perfil dos países pelos segmentos encontrados, e sugira para o grupo de pesquisa qual o melhor agrupamento de países para investigar indicadores de longevidade e doenças crônicas.

Case de Negócios: Uso de análise de *Cluster* no Sistema de Orçamentação

https://www.youtube.com/watch?v=V_gj04_xu3E

Fonte: YouTube - CESVI

Bateu? Nova ferramenta vai reduzir tempo do orçamento em 47%

O CESVI (Centro de Experimentação e Segurança Viária da Mapfre) anuncia o lançamento de sua **nova plataforma *Smart*** de orçamentos, que agiliza o processo de reparação usando a Inteligência Artificial.

https://carros.ig.com.br/2017-08-09/seguro-sinistro-orcamento.html

A partir das informações concebidas, o sistema traça a sugestão das peças a serem trocadas. **O usuário não precisa informar o estado de cada componente, pois o sistema é capaz de identificar as condições apenas com o laudo do mecânico**. Em outras palavras, a plataforma entende que uma colisão que abrange apenas o paralama e para-choque dificilmente comprometeria a suspensão, por exemplo. Um sinistro que afeta para-choque, paralamas, capô e portas dianteiras, por outro lado, traria um laudo totalmente diferente sobre as peças a serem substituídas e reaproveitadas a partir da gravidade da colisão.

Case de Negócios: Cesvi

http://www.cesvibrasil.com.br/Portal/Principal/Arquivos/Revista/Upload/_RC107_Simples.pdf

"Trabalhamos com um algoritmo proveniente da área de inteligência artificial que nos auxiliou no **agrupamento de intensidade de batida** com o intuito de **minimizar a variabilidade de peças** ofertadas em uma determinada versão de veículo"

Karin Tamura
Marketdata Solutions

Referências

- 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