Analytics e Inteligência Artificial Data Science

Tema da aula
Análise de Cluster

FIA BUSINESS SCHOOL lab.data

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 baseadas 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

BUSINESS SCHOOL lab.data

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.

+9.000 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 à estação de metrô (com estacionamento).

FIA BUSINESS SCHOOL lab.data

PROFA. DRA. ALESSANDRA DE ÁVILA MONTINI

Diretora do LABDATA-FIA, apaixonada por dados e pela arte de lecionar. Tem muito orgulho de ter criado na FIA cinco laboratórios para as aulas de Big Data e Inteligência Artificial. Possui mais de 20 anos de trajetória nas áreas de Data Mining, Big Data, Inteligência Artificial e Analytics. Cientista de dados com carreira realizada na Universidade de São Paulo. Graduada e mestra em Estatística Aplicada pelo IME-USP e doutora pela FEA-USP. Com muita dedicação chegou ao cargo de professora e pesquisadora na FEA-USP, ganhou mais de 30 prêmios de excelência acadêmica pela FEA-USP e mais de 30 prêmios de excelência acadêmica como professora dos cursos de MBA da FIA. Orienta alunos de mestrado e de doutorado na FEA-USP. Parecerista da FAPESP e colunista de grandes portais de tecnologia.

Currículo do Professor

PROF. ÂNGELO CHIODE, MSc

Bacharel, mestre e candidato ao PhD em Estatística (IME-USP), atua como professor de Estatística Aplicada para turmas de especialização, pós-graduação e MBA na FIA. Trabalha como consultor nas áreas de Analytics e Ciência de Dados há 13 anos, apoiando empresas na resolução de desafios de negócio nos contextos de finanças, adquirência, seguros, varejo, tecnologia, aviação, telecomunicações, entretenimento e saúde. Nos últimos 5 anos, tem atuado na gestão corporativa de times de Analytics, conduzindo projetos que envolviam análise estatística, modelagem preditiva e *machine learning*. É especializado em técnicas de visualização de dados e design da informação (Harvard) e foi indicado ao prêmio de Profissional do Ano na categoria Business Intelligence, em 2019, pela Associação Brasileira de Agentes Digitais (ABRADi).

Conteúdo Programático

DISCIPLINAS

- IA E TRANSFORMAÇÃO DIGITAL
- **ANALYTICS**
- INTELIGÊNCIA ARTIFICIAL: MACHINE LEARNING
- INTELIGÊNCIA ARTIFICIAL: DEEP LEARNING
- EMPREENDEDORISMO E INOVAÇÃO
- COMPORTAMENTO HUMANO E SOFT SKILLS

TEMAS: ANALYTICS E MACHINE LEARNING

- ANÁLISE EXPLORATÓRIA DE DADOS
- INFERÊNCIA ESTATÍSTICA
- TÉCNICAS DE PROJEÇÃO
- TÉCNICAS DE CLASSIFICAÇÃO
- TÓPICOS DE MODELAGEM
- **TÉCNICAS DE SEGMENTAÇÃO**
- TÓPICOS DE ANALYTICS
- MANIPULAÇÃO DE BASE DE DADOS
- AUTO ML

TEMAS: DEEP LEARNING

- REDES DENSAS
- REDES CONVOLUCIONAIS
- REDES RECORRENTES
- MODELOS GENERATIVOS

FERRAMENTAS

- **LINGUAGEM R**
- LINGUAGEM PYTHON
- DATABRICKS

FIA BUSINESS SCHOOL lab.data

Conteúdo da Aula

1. Introdução
2. Objetivo
3. Entendimento do Problema
4. Medidas de Distância
   i. Distância Euclidiana
   ii. *Simple Matching*
   iii. Distância de Gower
5. Padronização de Variáveis
   i. Método *Z-score*
   ii. Método *Range*
6. Algoritmo Hierárquico
   i. Dendrograma
   ii. Critérios de Ligação
7. Algoritmos de Partição
   i. K-Médias
   ii. K-Medoides
8. *Cases* Adicionais

Referências Bibliográficas

FIA BUSINESS SCHOOL lab.data

1. Introdução

Case: Encarteiramento de Clientes

Exemplo:
Criar um encarteiramento dos clientes de um banco para estabelecer níveis de atendimento diferenciados, de acordo com padrões de transacionalidade, posse de produtos e investimentos realizados.

Aplicação:
Segmento bancário

Case: Hábitos Alimentares

Exemplo:
Agrupar regiões do país com base na similaridade de hábitos alimentares.

Aplicação:
Áreas de saúde e nutrição

Case: Sequenciamento Genético

Exemplo:
Agrupar sequências de DNA com base em seu comprimento e frequência de ocorrência de certos padrões, a fim de compreender a estrutura e função dos genes.

Aplicação:
Área médica

Case: Comunicação Personalizada

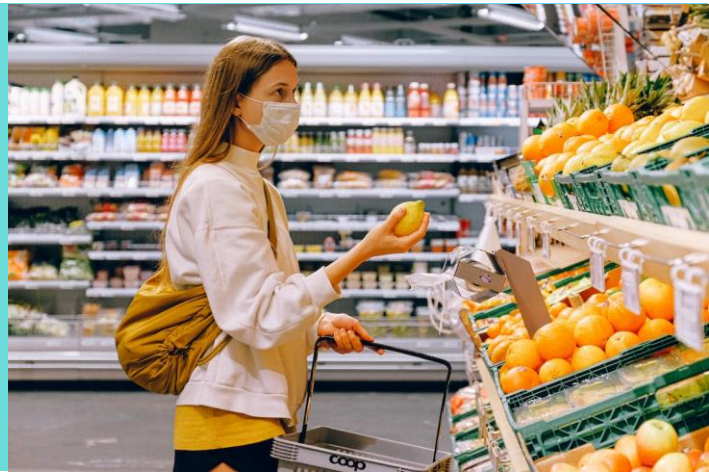
Exemplo:
Segmentar clientes de acordo com o seu perfil sociodemográfico, para implantar ações de comunicação personalizadas (marketing de relacionamento).

Aplicação:
Área de marketing e CRM

Case: Reconhecimento de Clientes

Exemplo:
Agrupar clientes de um varejo com base em recência, frequência e/ou valor de suas compras, a fim de estabelecer estratégias personalizadas de reconhecimento e relacionamento.

Aplicação:
Área de marketing e CRM

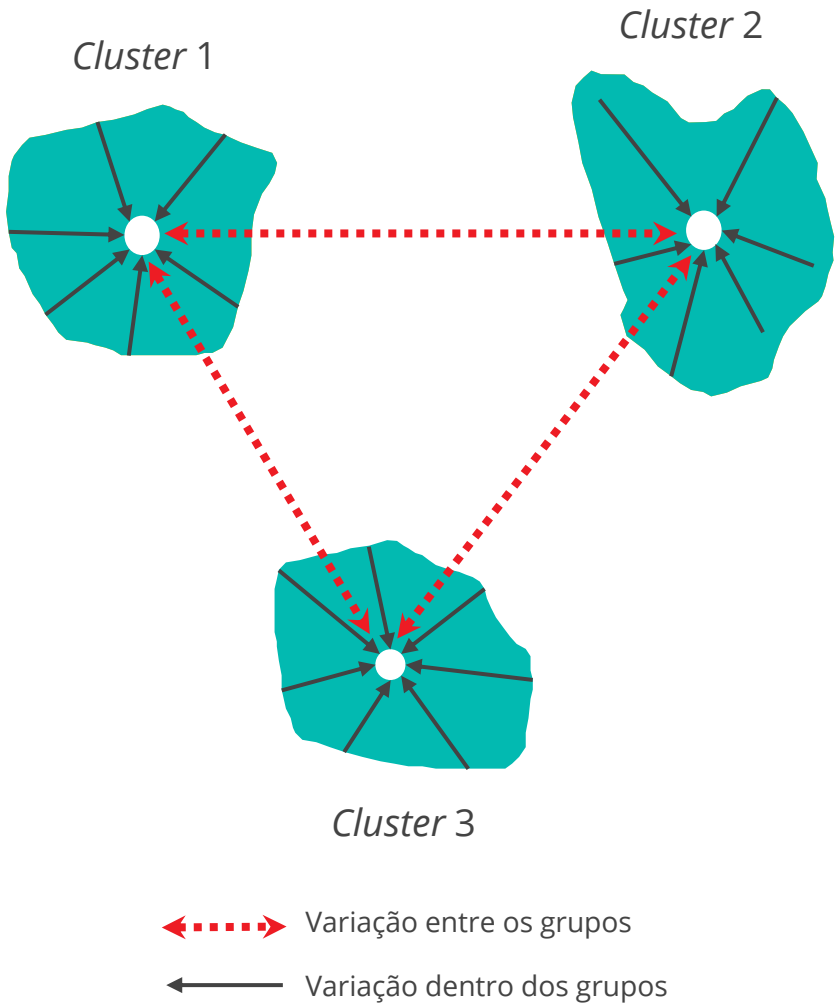


Introdução

- O que todos os *cases* anteriores possuem em comum?
- Qual a principal diferença entre esses *cases* e aqueles que resolvemos em aulas anteriores, por meio de técnicas de regressão linear e logística?

2. Objetivo

Objetivo



O objetivo da análise de *cluster* é **agrupar observações semelhantes** em uma base de dados.

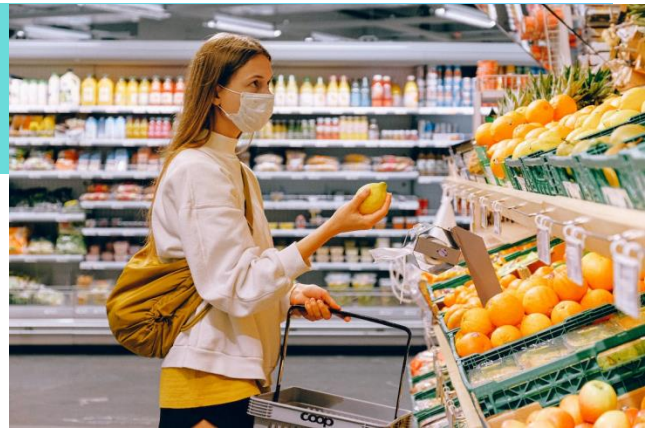
Tal agrupamento deve ocorrer de tal forma que, dentro de cada grupo, as observações sejam **homogêneas** (parecidas) entre si; e os grupos formados sejam **heterogêneos** (diferentes) entre si.

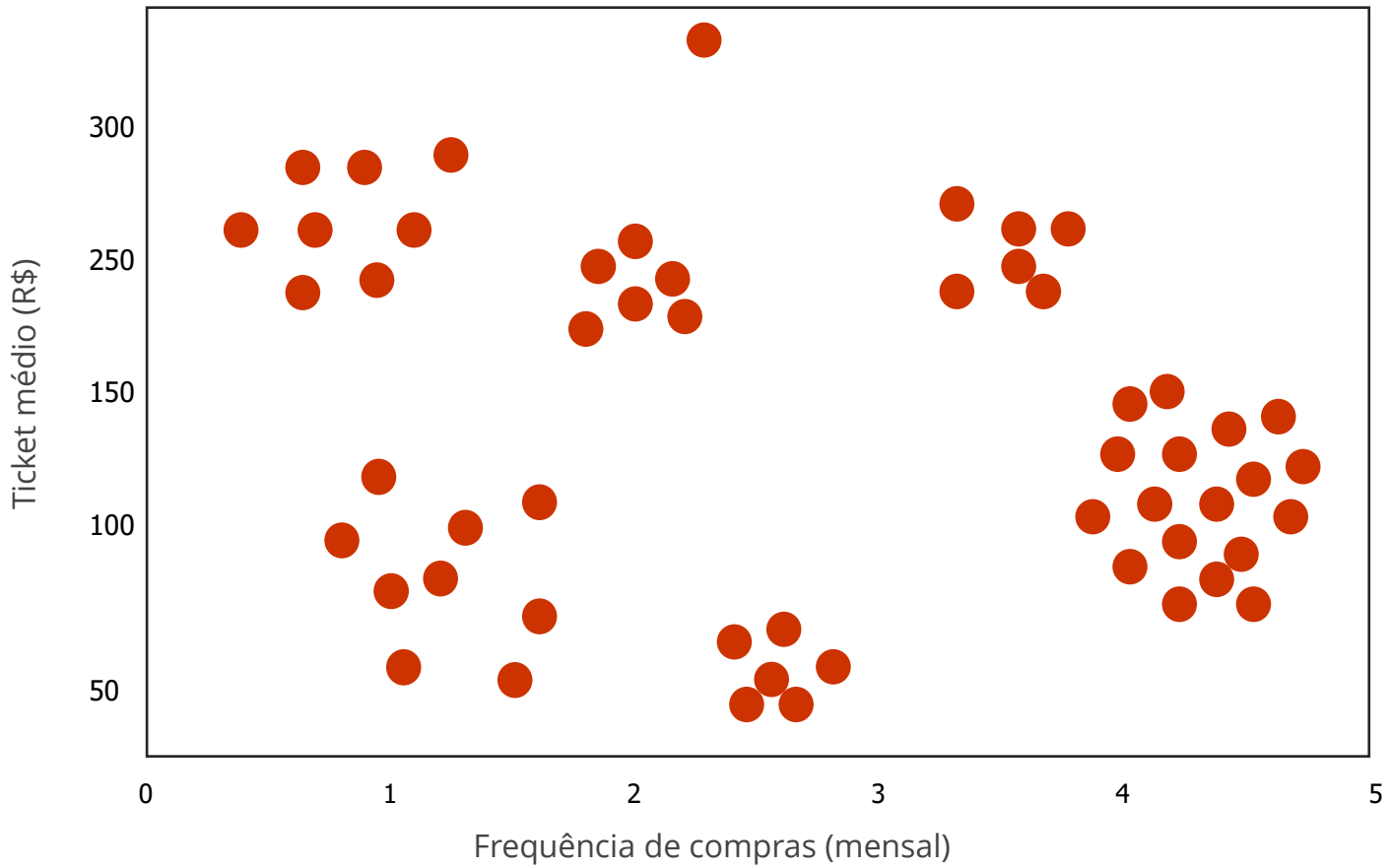
Ou seja, **dentro** de cada *cluster* (ou *grupo*), a variabilidade de características deve ser **mínima**; enquanto **entre** os *clusters*, a variabilidade deve ser **máxima**.

3. Entendimento do Problema

Case: Reconhecimento de Clientes

Criação de **4 *clusters*** de clientes de um varejo com base em **frequência** e **valor** de suas compras, a fim de estabelecer estratégias de reconhecimento e relacionamento.

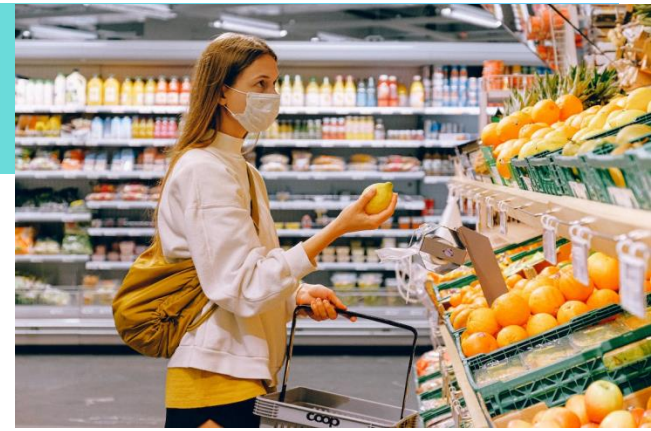


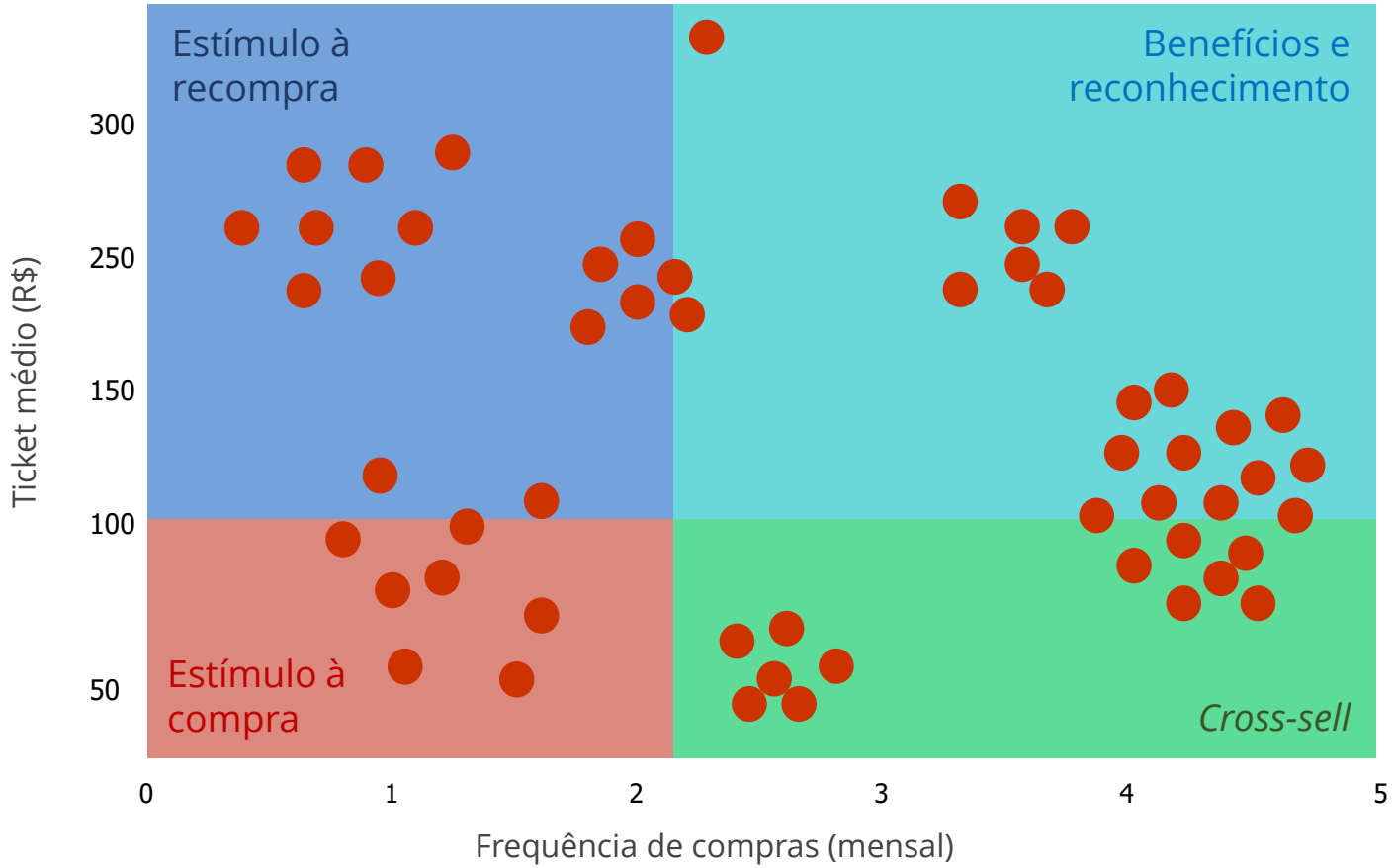


FIA BUSINESS SCHOOL lab.data

Case: Reconhecimento de Clientes

Criação de **4 *clusters*** de clientes de um varejo com base em **frequência** e **valor** de suas compras, a fim de estabelecer estratégias de reconhecimento e relacionamento.





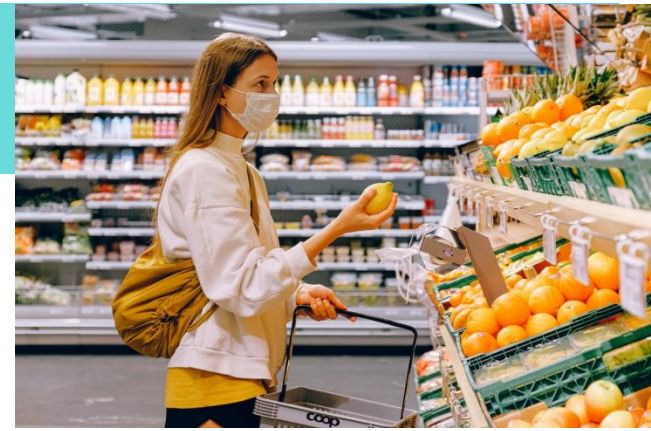
Foi proposta uma segmentação simples, dividida em **4 quadrantes**: acima/abaixo da média de cada um dos **dois eixos**.

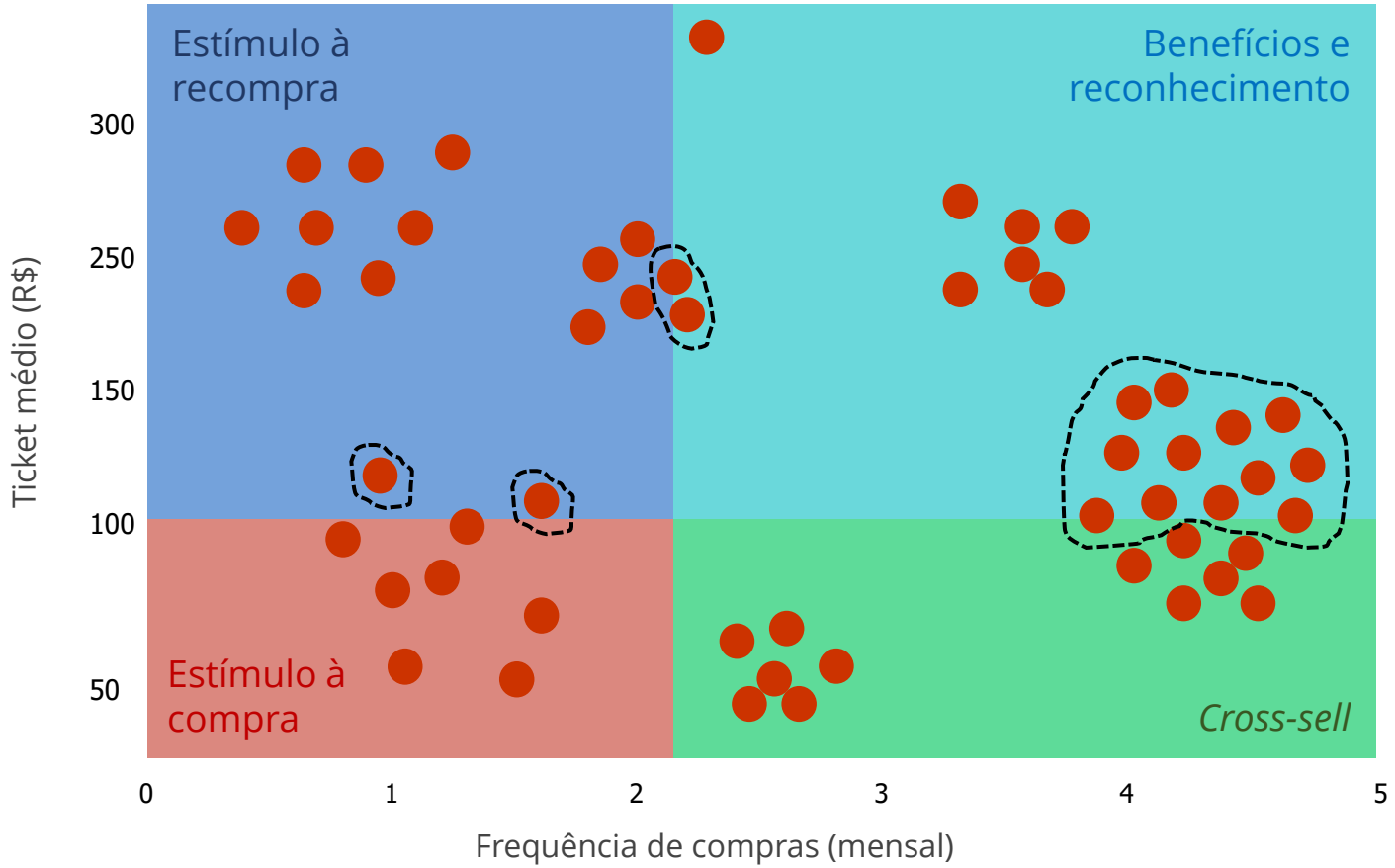
Essa segmentação de negócio parece adequada?

FIA BUSINESS SCHOOL lab.data

Case: Reconhecimento de Clientes

Criação de **4 *clusters*** de clientes de um varejo com base em **frequência** e **valor** de suas compras, a fim de estabelecer estratégias de reconhecimento e relacionamento.



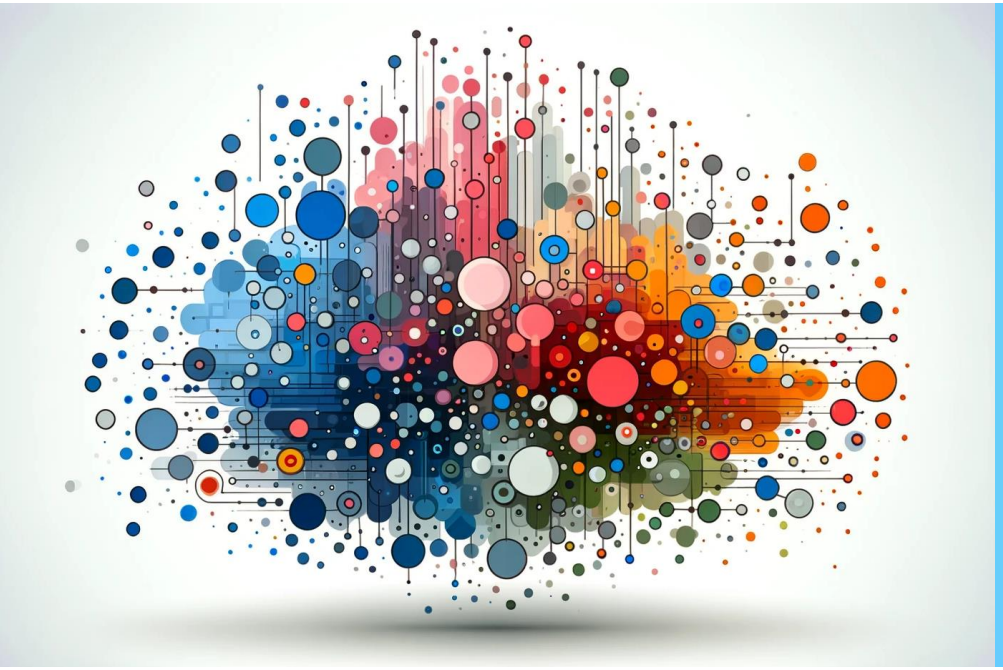


Uma segmentação baseada unicamente em **critérios de negócios** nem sempre fornece a melhor "regra" para agrupar observações semelhantes.

Para cumprir essa tarefa, precisaremos usar **métodos estatísticos.**

FIA BUSINESS SCHOOL lab.data

Identificação e Seleção de Variáveis



Duas perguntas naturais que surgem quando estamos estudando análise de *cluster*:

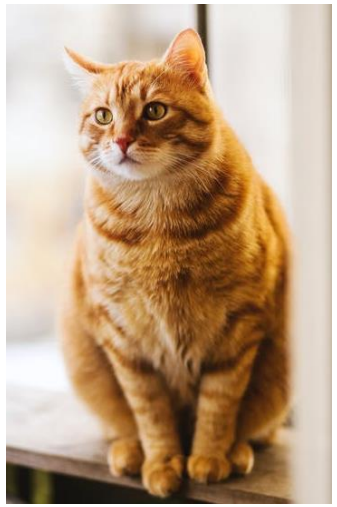
- Como devo **definir as variáveis** da minha análise?
- Será que existem métodos para selecionar as variáveis **mais importantes**, tal como na regressão?

Identificação e Seleção de Variáveis

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

?

?

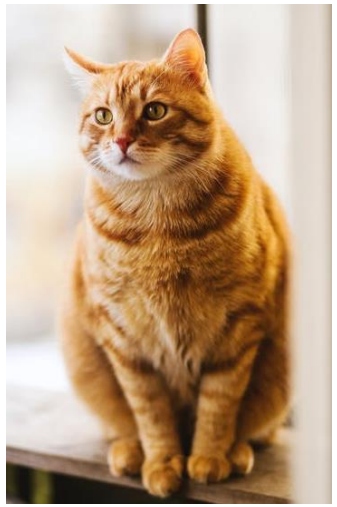


Identificação e Seleção de Variáveis

Depende! A semelhança entre as observações deve ser medida a depender da(s) **variável(is) de maior interesse.**

Porte

Elementos da face



Identificação e Seleção de Variáveis

Por se tratar de um método que não envolve variável resposta, **não há um critério estatístico de seleção de variáveis importantes** para uma análise de *cluster*.

Dessa forma, uma das etapas cruciais ao realizar essa análise consiste justamente em **definir as variáveis apropriadas**, de forma que reflitam o interesse do estudo.

Num contexto corporativo, os envolvidos nessa definição são tanto a área **técnica** de analytics/ciência de dados quanto a área de **negócios**. A primeira terá como missão transformar e/ou calcular variáveis de forma adequada, a fim de refletir os objetivos manifestados pela área de negócios.

4. Medidas de Distância

Definição de Medida de Distância

Uma vez definidas as variáveis para a nossa análise, é necessário estabelecer uma forma de **quantificar** a semelhança entre as observações. Isto pode ser realizado por meio das chamadas **medidas de distância**.

Estas medidas são nada mais do que **cálculos matemáticos** que nos ajudarão a mensurar o quão parecidas são duas observações entre si. O exemplo mais tradicional e intuitivo de medida é a **distância euclidiana**, que veremos a seguir.

FIA BUSINESS SCHOOL lab.data

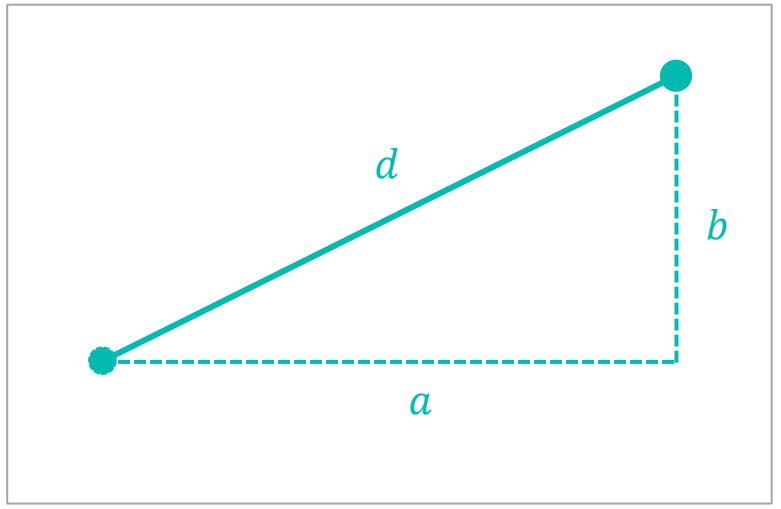
Distância Euclidiana

A **distância euclidiana** é uma medida de quão *similares/dissimilares* são duas observações em uma base de dados, caracterizadas a partir de 2 ou mais variáveis **quantitativas**.

Sua origem está respaldada no **Teorema de Pitágoras**, por meio do qual o comprimento do lado maior de um triângulo pode ser calculado a partir dos comprimentos dos lados menores.

Quanto **menor** o valor da distância euclidiana, **mais similares** são as duas observações entre si; e vice-versa.

Exemplo em 2 dimensões



Distância entre os dois pontos:

$$d = \sqrt{a^2 + b^2}$$

Case: Avaliação de Candidatos

Um recrutador de uma empresa de tecnologia deseja **segmentar 10 candidatos** a uma vaga, a fim de fornecer um diagnóstico resumido sobre seus perfis ao gestor contratante. Para isso, está considerando duas variáveis:

- tempo de experiência do candidato na área, em anos;
- quantidade de cursos ou especializações realizadas na área, após a graduação.

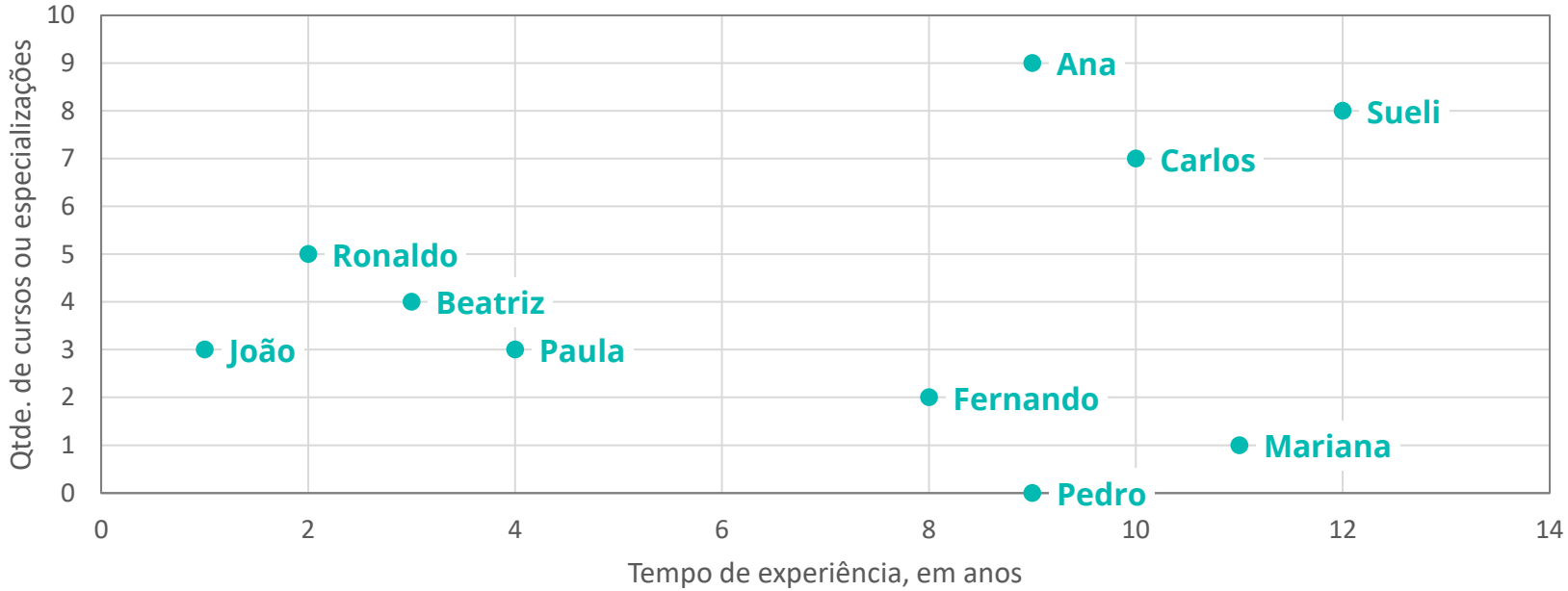
Candidato (a)	Tempo de experiência	Qtde. de cursos/especializações
Ana	9	9
Beatriz	3	4
Carlos	10	7
Fernando	8	2
João	1	3
Mariana	11	1
Paula	4	3
Pedro	9	0
Ronaldo	2	5
Sueli	12	8

Arquivo: Avaliacao_Candidatos (.txt)

Case: Avaliação de Candidatos

Um recrutador de uma empresa de tecnologia deseja **segmentar 10 candidatos** a uma vaga, a fim de fornecer um diagnóstico resumido sobre seus perfis ao gestor contratante. Para isso, está considerando duas variáveis:

- tempo de experiência do candidato na área, em anos;
- quantidade de cursos ou especializações realizadas na área, após a graduação.

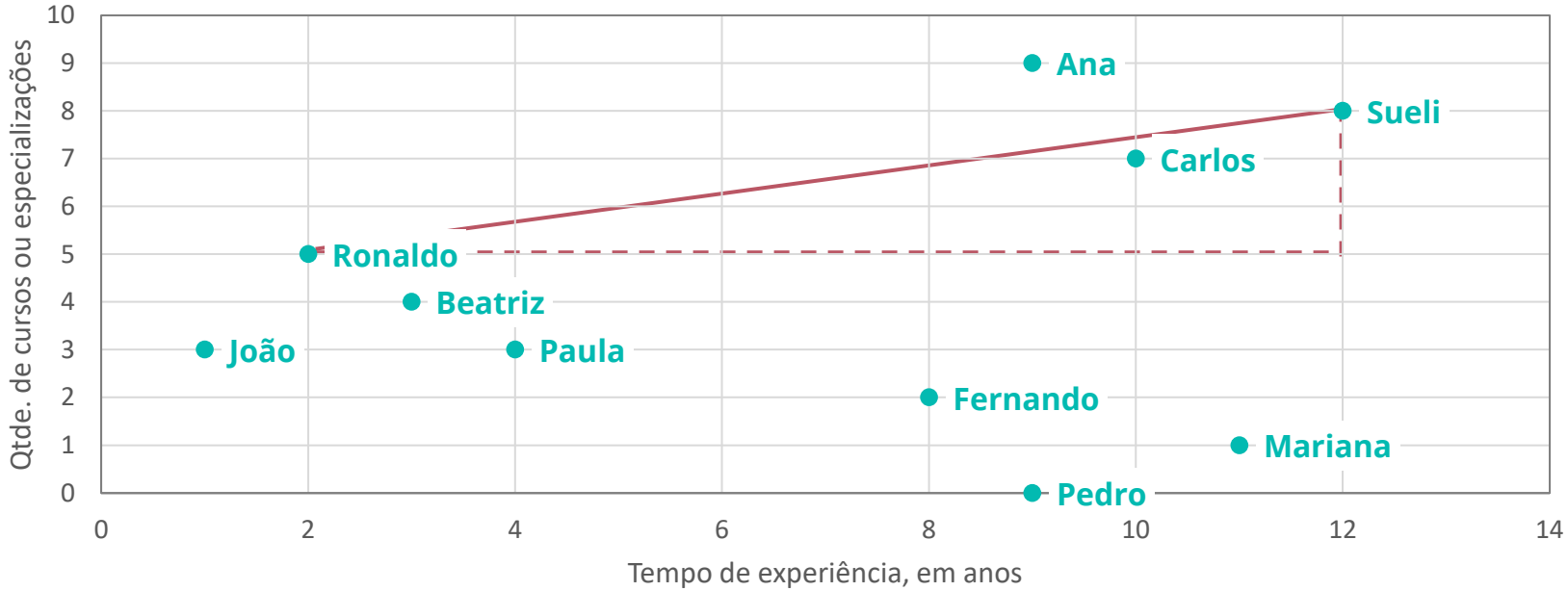


Arquivo: Avaliacao_Candidatos (.txt)

Case: Avaliação de Candidatos

Um recrutador de uma empresa de tecnologia deseja **segmentar 10 candidatos** a uma vaga, a fim de fornecer um diagnóstico resumido sobre seus perfis ao gestor contratante. Para isso, está considerando duas variáveis:

- tempo de experiência do candidato na área, em anos;
- quantidade de cursos ou especializações realizadas na área, após a graduação.



Distância euclidiana entre **Ronaldo** e **Sueli**:

$$d = \sqrt{(12-2)^2+(8-5)^2}$$

$$= \sqrt{10^2+3^2}$$

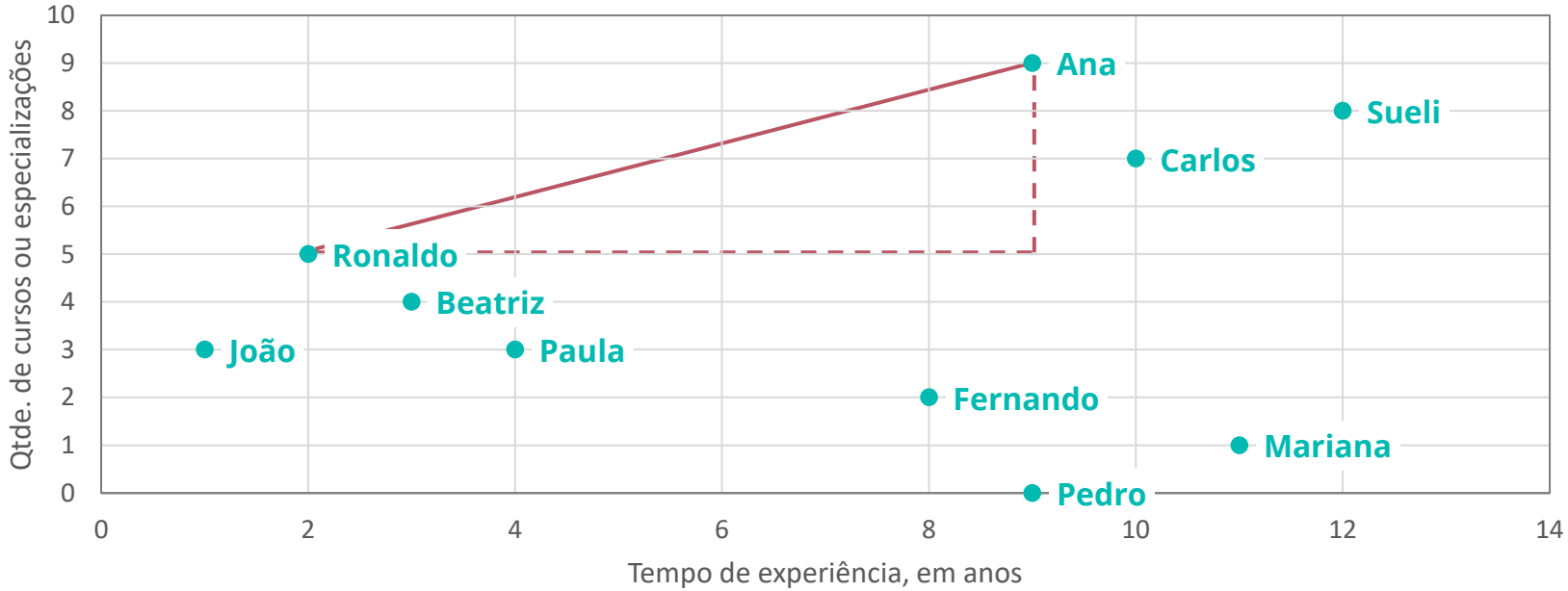
$$\approx 10{,}4$$

Arquivo: Avaliacao_Candidatos (.txt)

Case: Avaliação de Candidatos

Um recrutador de uma empresa de tecnologia deseja **segmentar 10 candidatos** a uma vaga, a fim de fornecer um diagnóstico resumido sobre seus perfis ao gestor contratante. Para isso, está considerando duas variáveis:

- tempo de experiência do candidato na área, em anos;
- quantidade de cursos ou especializações realizadas na área, após a graduação.



Distância euclidiana entre **Ronaldo** e **Ana**:

$$d = \sqrt{(9-2)^2+(9-5)^2}$$

$$= \sqrt{7^2+4^2}$$

$$\approx 8{,}1$$

Ronaldo *é mais parecido com* ***Ana*** *do que com* ***Sueli***

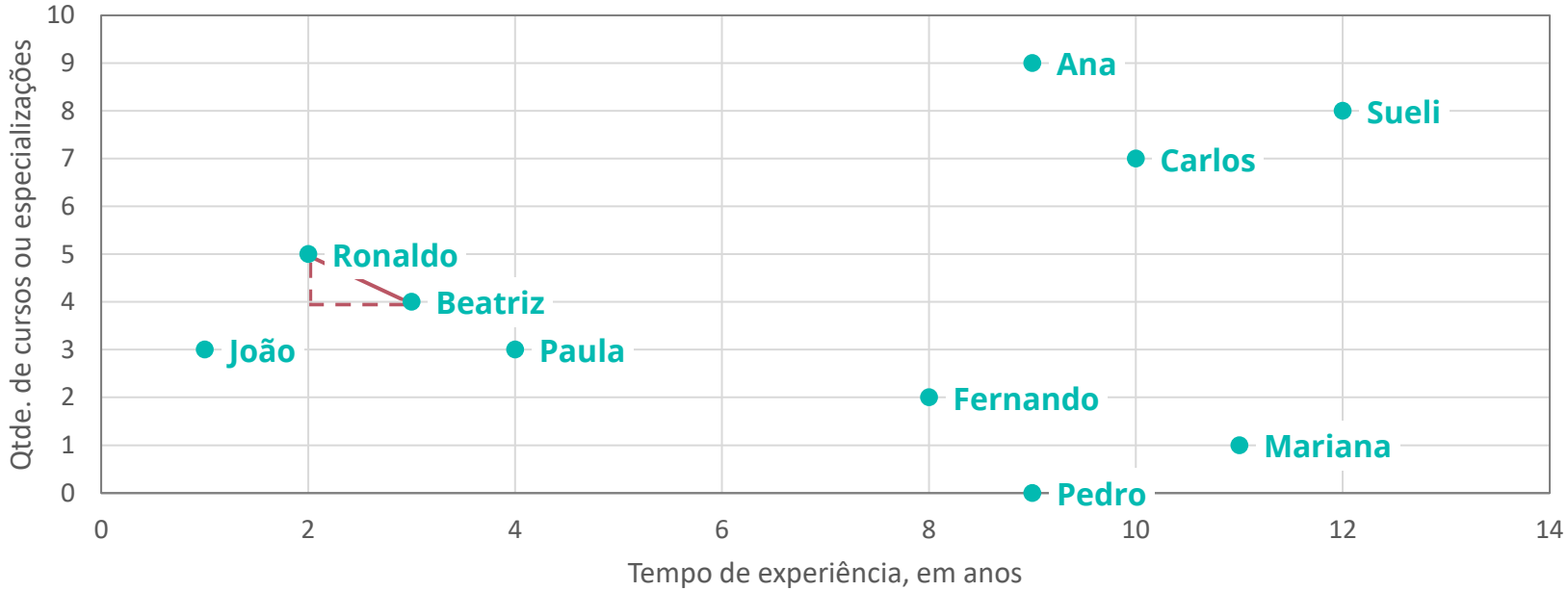
Arquivo: Avaliacao_Candidatos (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Avaliação de Candidatos

Um recrutador de uma empresa de tecnologia deseja **segmentar 10 candidatos** a uma vaga, a fim de fornecer um diagnóstico resumido sobre seus perfis ao gestor contratante. Para isso, está considerando duas variáveis:

- tempo de experiência do candidato na área, em anos;
- quantidade de cursos ou especializações realizadas na área, após a graduação.



Distância euclidiana entre **Ronaldo** e **Beatriz**:

$$d = \sqrt{(3-2)^2+(5-4)^2}$$

$$= \sqrt{1^2+1^2}$$

$$\approx 1{,}4$$

Ronaldo** é mais parecido com **Beatriz** do que com **Ana** e **Sueli

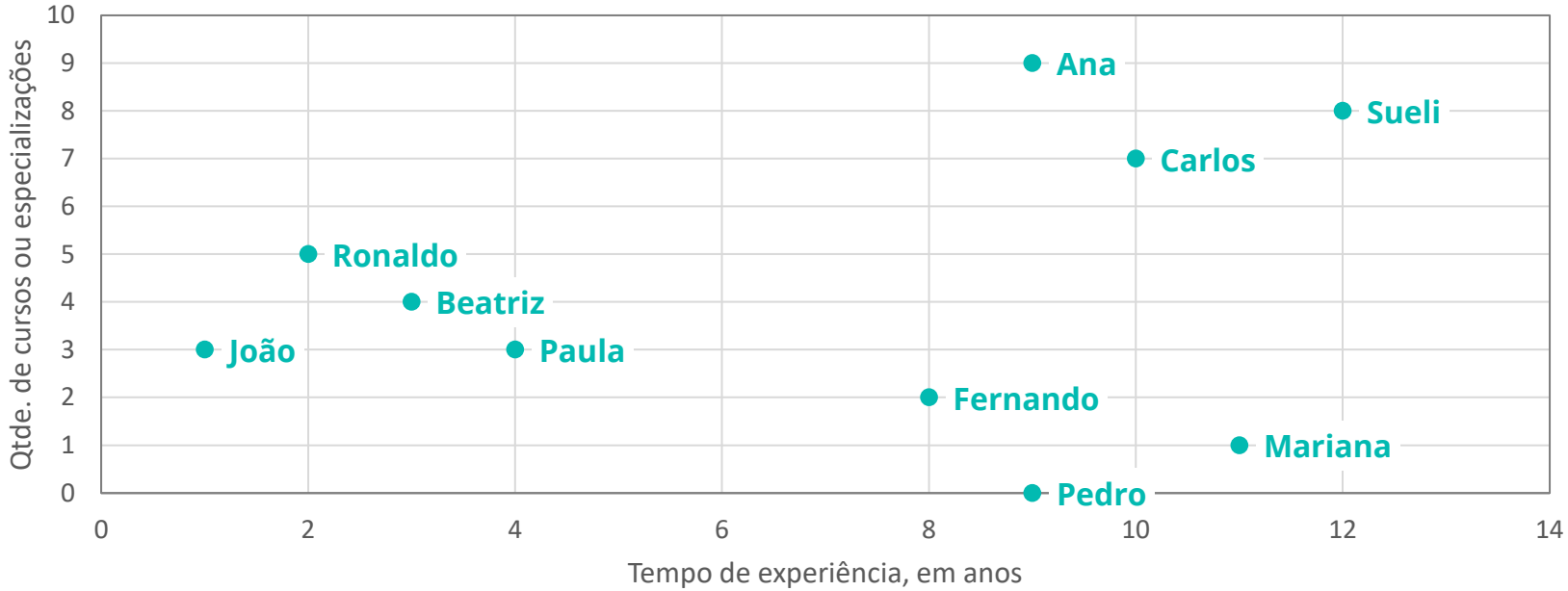
Arquivo: Avaliacao_Candidatos (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Avaliação de Candidatos

Um recrutador de uma empresa de tecnologia deseja **segmentar 10 candidatos** a uma vaga, a fim de fornecer um diagnóstico resumido sobre seus perfis ao gestor contratante. Para isso, está considerando duas variáveis:

- tempo de experiência do candidato na área, em anos;
- quantidade de cursos ou especializações realizadas na área, após a graduação.



Com base na análise visual das distâncias euclidianas, quantos grupos de candidatos parecem existir?

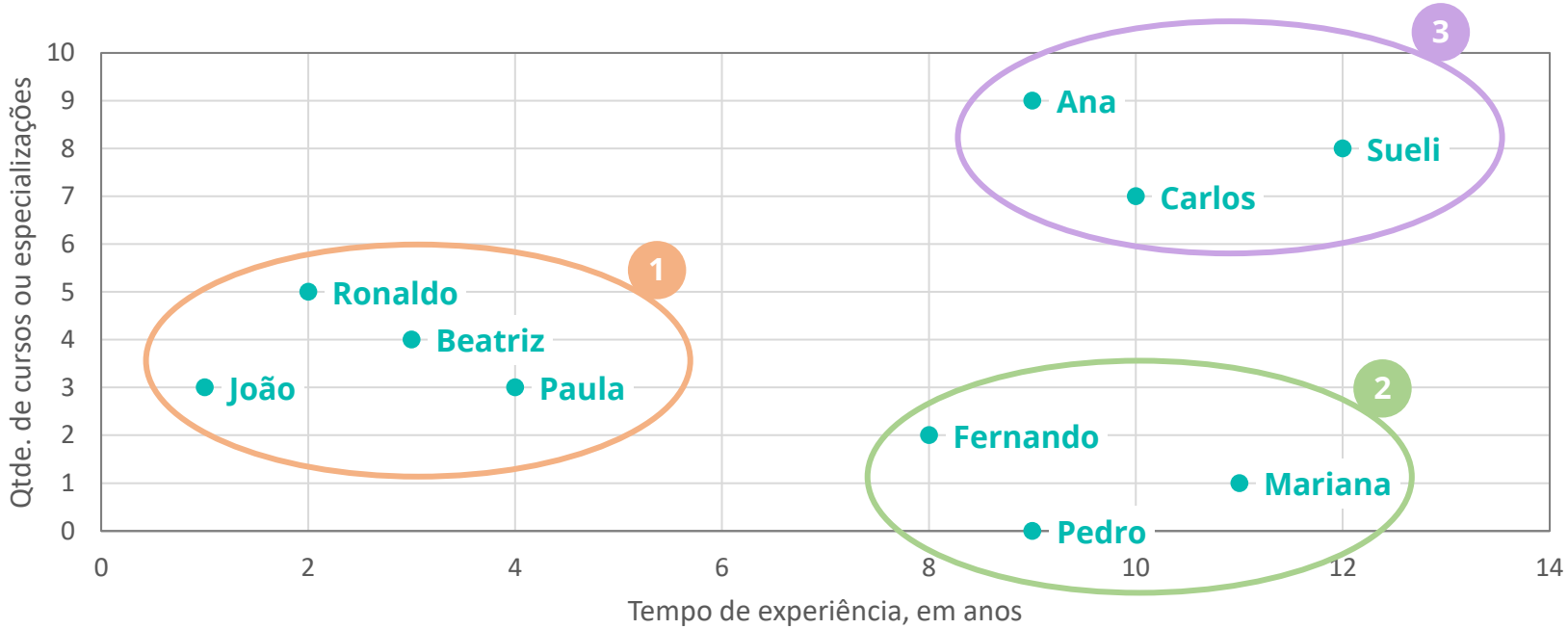
Arquivo: Avaliacao_Candidatos (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Avaliação de Candidatos

Um recrutador de uma empresa de tecnologia deseja **segmentar 10 candidatos** a uma vaga, a fim de fornecer um diagnóstico resumido sobre seus perfis ao gestor contratante. Para isso, está considerando duas variáveis:

- tempo de experiência do candidato na área, em anos;
- quantidade de cursos ou especializações realizadas na área, após a graduação.



Com base na análise visual das distâncias euclidianas, quantos grupos de candidatos parecem existir?

Arquivo: Avaliacao_Candidatos (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Avaliação de Candidatos

Consolidando as medidas de distância euclidiana de todos os pares de candidatos, podemos obter uma **matriz de distâncias**:

Candidato	Ana	Beatriz	Carlos	Fernando	João	Mariana	Paula	Pedro	Ronaldo	Sueli
Ana		7,8	2,2	7,1	10,0	8,2	7,8	9,0	8,1	3,2
Beatriz			7,6	5,4	2,2	8,5	1,4	7,2	1,4	9,8
Carlos				5,4	9,8	6,1	7,2	7,1	8,2	2,2
Fernando					7,1	3,2	4,1	2,2	6,7	7,2
João						10,2	3,0	8,5	2,2	12,1
Mariana							7,3	2,2	9,8	7,1
Paula								5,8	2,8	9,4
Pedro									8,6	8,5
Ronaldo										10,4
Sueli										

Arquivo: Avaliacao_Candidatos (.txt)

Case: Avaliação de Candidatos

Consolidando as medidas de distância euclidiana de todos os pares de candidatos, podemos obter uma **matriz de distâncias**:

Candidato	Ana	Beatriz	Carlos	Fernando	João	Mariana	Paula	Pedro	Ronaldo	Sueli
Ana		7,8	2,2	7,1	10,0	8,2	7,8	9,0	8,1	3,2
Beatriz			7,6	5,4	2,2	8,5	1,4	7,2	1,4	9,8
Carlos				5,4	9,8	6,1	7,2	7,1	8,2	2,2
Fernando					7,1	3,2	4,1	2,2	6,7	7,2
João						10,2	3,0	8,5	2,2	12,1
Mariana							7,3	2,2	9,8	7,1
Paula								5,8	2,8	9,4
Pedro									8,6	8,5
Ronaldo										10,4
Sueli										

- Por que os valores da diagonal (em fundo azul) foram omitidos?
- Por que os valores abaixo da diagonal também foram omitidos?

Arquivo: Avaliacao_Candidatos (.txt)

Case: Avaliação de Candidatos

Consolidando as medidas de distância euclidiana de todos os pares de candidatos, podemos obter uma **matriz de distâncias**:

Candidato	Ana	Beatriz	Carlos	Fernando	João	Mariana	Paula	Pedro	Ronaldo	Sueli
Ana		7,8	2,2	7,1	10,0	8,2	7,8	9,0	8,1	3,2
Beatriz			7,6	5,4	2,2	8,5	1,4	7,2	1,4	9,8
Carlos				5,4	9,8	6,1	7,2	7,1	8,2	2,2
Fernando					7,1	3,2	4,1	2,2	6,7	7,2
João						10,2	3,0	8,5	2,2	12,1
Mariana							7,3	2,2	9,8	7,1
Paula								5,8	2,8	9,4
Pedro									8,6	8,5
Ronaldo										10,4
Sueli										

Distâncias entre os candidatos do **grupo 1** (identificado visualmente)

Arquivo: Avaliacao_Candidatos (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Avaliação de Candidatos

Consolidando as medidas de distância euclidiana de todos os pares de candidatos, podemos obter uma **matriz de distâncias**:

Candidato	Ana	Beatriz	Carlos	Fernando	João	Mariana	Paula	Pedro	Ronaldo	Sueli
Ana		7,8	2,2	7,1	10,0	8,2	7,8	9,0	8,1	3,2
Beatriz			7,6	5,4	2,2	8,5	1,4	7,2	1,4	9,8
Carlos				5,4	9,8	6,1	7,2	7,1	8,2	2,2
Fernando					7,1	3,2	4,1	2,2	6,7	7,2
João						10,2	3,0	8,5	2,2	12,1
Mariana							7,3	2,2	9,8	7,1
Paula								5,8	2,8	9,4
Pedro									8,6	8,5
Ronaldo										10,4
Sueli										

Distâncias entre os candidatos do **grupo 2** (identificado visualmente)

Arquivo: Avaliacao_Candidatos (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Avaliação de Candidatos

Consolidando as medidas de distância euclidiana de todos os pares de candidatos, podemos obter uma **matriz de distâncias**:

Candidato	Ana	Beatriz	Carlos	Fernando	João	Mariana	Paula	Pedro	Ronaldo	Sueli
Ana		7,8	2,2	7,1	10,0	8,2	7,8	9,0	8,1	3,2
Beatriz			7,6	5,4	2,2	8,5	1,4	7,2	1,4	9,8
Carlos				5,4	9,8	6,1	7,2	7,1	8,2	2,2
Fernando					7,1	3,2	4,1	2,2	6,7	7,2
João						10,2	3,0	8,5	2,2	12,1
Mariana							7,3	2,2	9,8	7,1
Paula								5,8	2,8	9,4
Pedro									8,6	8,5
Ronaldo										10,4
Sueli										

Distâncias entre os candidatos do **grupo 3** (identificado visualmente)

Arquivo: Avaliacao_Candidatos (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Avaliação de Candidatos

Consolidando as medidas de distância euclidiana de todos os pares de candidatos, podemos obter uma **matriz de distâncias**:

Candidato	Ana	Beatriz	Carlos	Fernando	João	Mariana	Paula	Pedro	Ronaldo	Sueli
Ana		7,8	2,2	7,1	10,0	8,2	7,8	9,0	8,1	3,2
Beatriz			7,6	5,4	2,2	8,5	1,4	7,2	1,4	9,8
Carlos				5,4	9,8	6,1	7,2	7,1	8,2	2,2
Fernando					7,1	3,2	4,1	2,2	6,7	7,2
João						10,2	3,0	8,5	2,2	12,1
Mariana							7,3	2,2	9,8	7,1
Paula								5,8	2,8	9,4
Pedro									8,6	8,5
Ronaldo										10,4
Sueli										

Distâncias entre candidatos de **grupos distintos** (identificados visualmente)

Arquivo: Avaliacao_Candidatos (.txt)

FIA BUSINESS SCHOOL lab.data

Simple Matching

O ***simple matching*** é uma medida de quão *similares/dissimilares* são duas observações em uma base de dados, caracterizadas a partir de 2 ou mais variáveis **qualitativas**.

Basicamente, o *simple matching* corresponde a quantidade de variáveis em que houve correspondência (*"match"*) da categoria observada entre os dois elementos comparados.

Quanto **maior** o valor do *simple matching*, **mais similares** são as duas observações entre si; e vice-versa.

Exemplo em 3 dimensões

Obs.	Gênero	Faixa etária	Possui produto?
Obs. 1	Masculino	18 a 25	Sim
Obs. 2	Feminino	18 a 25	Sim
Obs. 3	Feminino	45 a 55	Não

Matching entre observações **1** e **2**:

$$m = 2$$

Matching entre observações **1** e **3**:

$$m = 0$$

Matching entre observações **2** e **3**:

$$m = 1$$

Case: Avaliação de Candidatos

Voltando ao *case* de avaliação de candidatos, suponhamos que o recrutador decidiu segmentar os candidatos sob outra ótica, agora considerando as duas seguintes variáveis:

- área de formação do candidato;
- nível hierárquico exercido no emprego atual: pleno ou sênior.

Candidato (a)	Área de formação	Nível hierárquico atual
Ana	Engenharia	Sênior
Beatriz	Ciência da Computação	Pleno
Carlos	Análise de Sistemas	Sênior
Fernando	Ciência da Computação	Sênior
João	Engenharia	Pleno
Mariana	Ciência da Computação	Pleno
Paula	Análise de Sistemas	Sênior
Pedro	Análise de Sistemas	Pleno
Ronaldo	Ciência da Computação	Pleno
Sueli	Engenharia	Sênior

Arquivo: Avaliacao_Candidatos (.txt)

Case: Avaliação de Candidatos

Voltando ao *case* de avaliação de candidatos, suponhamos que o recrutador decidiu segmentar os candidatos sob outra ótica, agora considerando as duas seguintes variáveis:

- área de formação do candidato;
- nível hierárquico exercido no emprego atual: pleno ou sênior.

Candidato (a)	Área de formação	Nível hierárquico atual
Ana	Engenharia	Sênior
Beatriz	Ciência da Computação	Pleno
Carlos	Análise de Sistemas	Sênior
Fernando	Ciência da Computação	Sênior
João	Engenharia	Pleno
Mariana	Ciência da Computação	Pleno
Paula	Análise de Sistemas	Sênior
Pedro	Análise de Sistemas	Pleno
Ronaldo	Ciência da Computação	Pleno
Sueli	Engenharia	Sênior

Beatriz, Mariana e Ronaldo são totalmente similares entre si segundo o *simple matching* *(d = 2)*.

Arquivo: Avaliacao_Candidatos (.txt)

Case: Avaliação de Candidatos

Voltando ao *case* de avaliação de candidatos, suponhamos que o recrutador decidiu segmentar os candidatos sob outra ótica, agora considerando as duas seguintes variáveis:

- área de formação do candidato;
- nível hierárquico exercido no emprego atual: pleno ou sênior.

Candidato (a)	Área de formação	Nível hierárquico atual
Ana	Engenharia	Sênior
Beatriz	Ciência da Computação	Pleno
Carlos	Análise de Sistemas	Sênior
Fernando	Ciência da Computação	Sênior
João	Engenharia	Pleno
Mariana	Ciência da Computação	Pleno
Paula	Análise de Sistemas	Sênior
Pedro	Análise de Sistemas	Pleno
Ronaldo	Ciência da Computação	Pleno
Sueli	Engenharia	Sênior

Já **Carlos** e **Pedro** não são totalmente similares entre si, por conta de seus níveis hierárquicos *(d = 1)*.

Arquivo: Avaliacao_Candidatos (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Avaliação de Candidatos

Voltando ao *case* de avaliação de candidatos, suponhamos que o recrutador decidiu segmentar os candidatos sob outra ótica, agora considerando as duas seguintes variáveis:

- área de formação do candidato;
- nível hierárquico exercido no emprego atual: pleno ou sênior.

Candidato (a)	Área de formação	Nível hierárquico atual
Ana	Engenharia	Sênior
Beatriz	Ciência da Computação	Pleno
Carlos	Análise de Sistemas	Sênior
Fernando	Ciência da Computação	Sênior
João	Engenharia	Pleno
Mariana	Ciência da Computação	Pleno
Paula	Análise de Sistemas	Sênior
Pedro	Análise de Sistemas	Pleno
Ronaldo	Ciência da Computação	Pleno
Sueli	Engenharia	Sênior

Por fim, **João** e **Paula** são totalmente distintos entre si segundo o *simple matching* ($d = 0$).

Arquivo: Avaliacao_Candidatos (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Avaliação de Candidatos

De forma análoga à distância euclidiana, é possível consolidar os valores de *simple matching* de todos os pares de candidatos por meio de uma **matriz**:

Candidato	Ana	Beatriz	Carlos	Fernando	João	Mariana	Paula	Pedro	Ronaldo	Sueli
Ana		0	1	1	1	0	1	0	0	2
Beatriz			0	1	1	2	0	1	2	0
Carlos				1	0	0	2	1	0	1
Fernando					0	1	1	0	1	1
João						1	0	1	1	1
Mariana							0	1	2	0
Paula								1	0	1
Pedro									1	0
Ronaldo										0
Sueli										

Arquivo: Avaliacao_Candidatos (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Avaliação de Candidatos

De forma análoga à distância euclidiana, é possível consolidar os valores de *simple matching* de todos os pares de candidatos por meio de uma **matriz**:

Candidato	Ana	Beatriz	Carlos	Fernando	João	Mariana	Paula	Pedro	Ronaldo	Sueli
Ana		0	1	1	1	0	1	0	0	2
Beatriz			0	1	1	2	0	1	2	0
Carlos				1	0	0	2	1	0	1
Fernando					0	1	1	0	1	1
João						1	0	1	1	1
Mariana							0	1	2	0
Paula								1	0	1
Pedro									1	0
Ronaldo										0
Sueli										

Note que, quando há poucas variáveis categóricas, existe pouca heterogeneidade para ser mensurada por meio do *simple matching*. Esse cenário fica mais interessante quando se tem um número maior de variáveis.

Arquivo: Avaliacao_Candidatos (.txt)

Distância de Gower

A **distância de Gower** é uma medida de quão *similares/dissimilares* são duas observações em uma base de dados, que podem ser caracterizadas tanto a partir de variáveis **quantitativas** quanto **qualitativas**.

O cálculo da distância de Gower é um pouco mais complexo* que o das medidas que discutimos anteriormente. Por outro lado, ela é **padronizada** no intervalo de 0 a 1, o que facilita a sua interpretação.

Quanto **mais próximo de 0** o valor da distância de Gower, **mais similares** são as duas observações entre si; e vice-versa.

Similaridade parcial entre os indivíduos *i* e *j* para uma variável **quantitativa** X:

$$s_{ij} = 1 - \frac{|X_i - X_j|}{\max(X) - \min(X)}$$

Similaridade parcial entre os indivíduos *i* e *j* para uma variável **qualitativa** Y:

$$s_{ij} = \begin{cases} 1 \ \ se \ Y_i = Y_j \\ 0 \ \ se \ Y_i \neq Y_j \end{cases}$$

Distância de Gower para os indivíduos *i* e *j*:

$$D_{ij} = 1 - \frac{1}{p}\sum_{k=1}^{p} s_{ij_{(k)}}$$

** Referência adicional: https://towardsdatascience.com/clustering-on-mixed-data-types-5fe226f9d9ca*

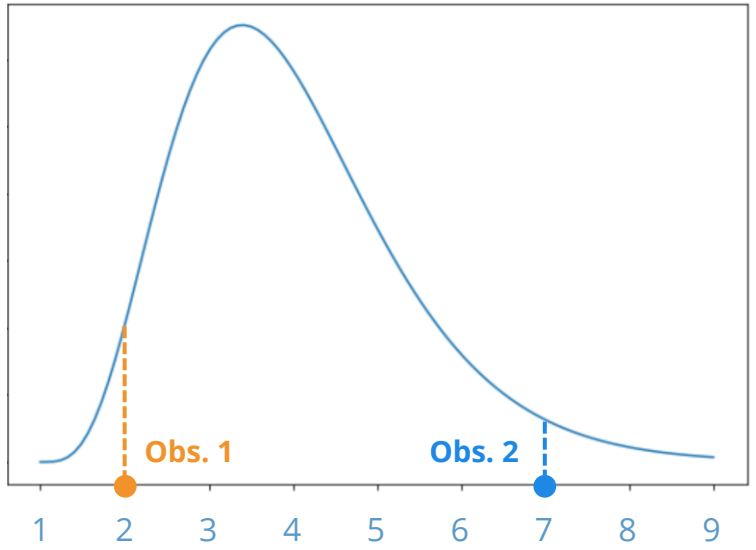
FIA BUSINESS SCHOOL lab.data

5. Padronização de Variáveis

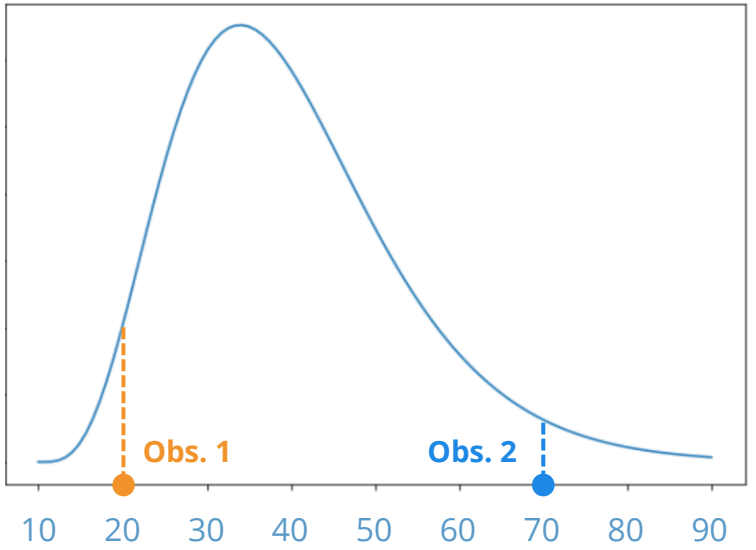
Racional da Necessidade de Padronização

Quando calculamos distâncias baseadas em variáveis quantitativas, variáveis com **escalas diferentes** tendem a exercer um **pesos diferentes** nos cálculos.

Exemplo: Variável X_1



Exemplo: Variável X_2



Distância euclidiana entre obs. 1 e obs. 2

$$D = \sqrt{(7-2)^2 + (70-20)^2}$$

$$= \sqrt{5^2 + 50^2}$$

$$= \sqrt{25 + 2.500}$$

$$= \mathbf{50,2}$$

Distância extremamente influenciada pela **variável X_2**, cuja escala é mais elevada

Método *Z-score*

Para evitar esse problema, deve-se realizar uma **padronização** das variáveis quantitativas.

O método mais usual de padronização é o ***z-score***, por meio do qual os valores de cada variável quantitativa são recalculados subtraindo a **média** da respectiva variável e dividindo pelo seu **desvio padrão**.

Ou seja, para cada valor x_i de uma variável X, temos:

$$x_i\ padronizado = \frac{x_i - \text{média}\ (X)}{\text{d. p.}\ (X)}$$

Com essa padronização, o novo conjunto de valores da variável X padronizada apresentará **média = 0** e **desvio padrão = 1**.

Método *Range*

Um método alternativo para padronizar as variáveis quantitativas é o ***range***.

Esse método relativiza os valores de uma variável a partir de sua **amplitude**, padronizando-os de forma que variem necessariamente numa escala de **0 a 1**.

Para cada valor x_i de uma variável X, temos:

$$x_i\ padronizado = \frac{x_i - \text{mín}\ (X)}{\text{máx}\ (X) - \text{mín}\ (X)}$$

Apesar de bastante intuitivo, este método pode distorcer o comportamento das variáveis caso elas possuem forte **assimetria** ou valores ***outliers*** acentuados. Estes fatores podem causar um "achatamento" da variável padronizada em um intervalo de valores mais restrito.

Case: Avaliação de Candidatos

Padronizando as duas variáveis quantitativas no *case* de avaliação de candidatos, por meio do ***z-score***.

Candidato (a)	Tempo de experiência	Tempo de experiência **padronizado**	Qtde. de cursos / especializações	Qtde. de cursos / especializações **padronizada**
Ana	9	(9 – 6,9) / 4,0 = **0,5**	9	(9 – 4,2) / 3,0 = **1,6**
Beatriz	3	(3 – 6,9) / 4,0 = **-1,0**	4	(4 – 4,2) / 3,0 = **-0,1**
Carlos	10	(10 – 6,9) / 4,0 = **0,8**	7	(7 – 4,2) / 3,0 = **0,9**
Fernando	8	(8 – 6,9) / 4,0 = **0,3**	2	(2 – 4,2) / 3,0 = **-0,7**
João	1	(1 – 6,9) / 4,0 = **-1,5**	3	(3 – 4,2) / 3,0 = **-0,4**
Mariana	11	(11 – 6,9) / 4,0 = **1,0**	1	(1 – 4,2) / 3,0 = **-1,1**
Paula	4	(4 – 6,9) / 4,0 = **-0,7**	3	(3 – 4,2) / 3,0 = **-0,4**
Pedro	9	(9 – 6,9) / 4,0 = **0,5**	0	(0 – 4,2) / 3,0 = **-1,4**
Ronaldo	2	(2 – 6,9) / 4,0 = **-1,2**	5	(5 – 4,2) / 3,0 = **0,3**
Sueli	12	(12 – 6,9) / 4,0 = **1,3**	8	(8 – 4,2) / 3,0 = **1,3**
Média	**6,9**	**0**	**4,2**	**0**
Desvio padrão	**4,0**	**1**	**3,0**	**1**

Arquivo: Avaliacao_Candidatos (.txt)

Case: Avaliação de Candidatos

Matriz de distâncias euclidianas **atualizada**, para as variáveis padronizadas:

Candidato	Ana	Beatriz	Carlos	Fernando	João	Mariana	Paula	Pedro	Ronaldo	Sueli
Ana		2,2	0,7	2,3	2,8	2,7	2,4	3,0	2,2	0,8
Beatriz			2,0	1,4	0,6	2,2	0,4	2,0	0,4	2,6
Carlos				1,7	2,6	2,0	2,0	2,3	2,1	0,6
Fernando					1,8	0,8	1,1	0,7	1,8	2,2
João						2,6	0,7	2,2	0,7	3,2
Mariana							1,9	0,6	2,6	2,3
Paula								1,6	0,8	2,6
Pedro									2,4	2,8
Ronaldo										2,7
Sueli										

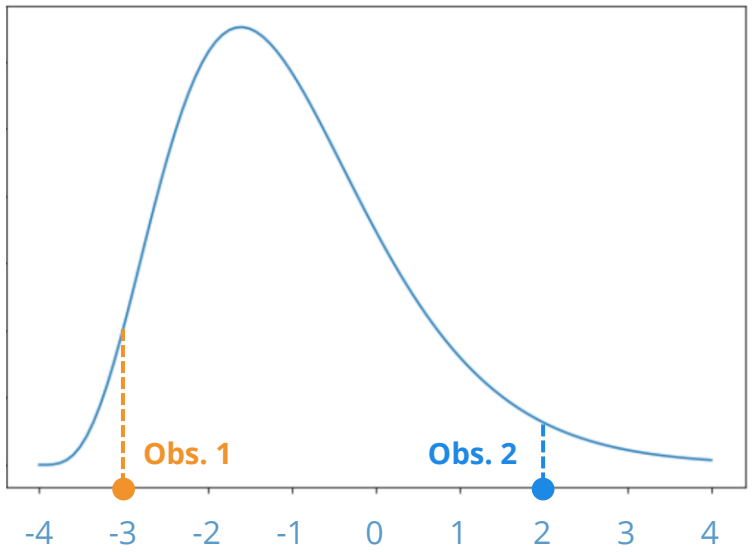
Arquivo: Avaliacao_Candidatos (.txt)

Considerações sobre a Padronização

- Mesmo que as variáveis da *clusterização* possuam escalas aproximadamente iguais, a boa prática consiste em **sempre padronizar** as variáveis quantitativas, a fim de que não fiquemos "reféns" das diferenças de escala ao calcular a medida de distância.
- Caso se tenha interesse explícito em atribuir **maior ou menor peso** para alguma(s) das variáveis, isso pode ser realizado após a padronização, por meio da **multiplicação** por algum fator arbitrário.
- Ao multiplicar uma variável por algum valor **maior do que 1**, aumentamos a sua influência na segmentação. Já se multiplicarmos por algum valor **maior que 0 e menor que 1**, diminuímos a sua influência.

Ampliando ou Reduzindo a Escala

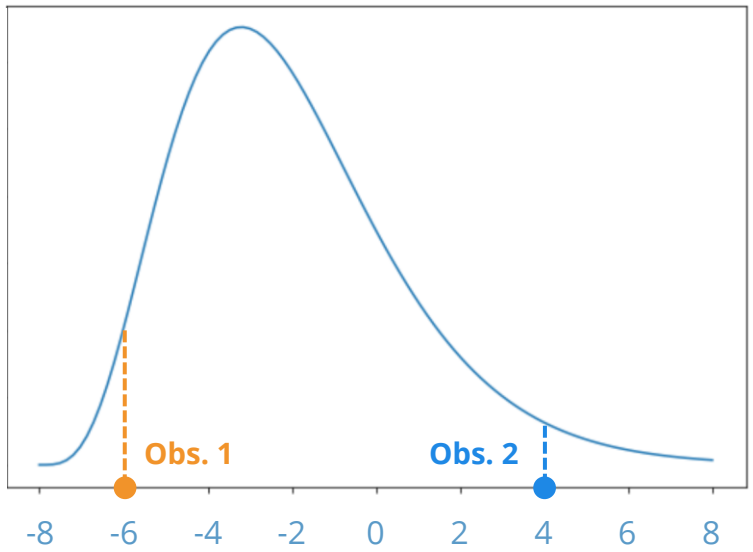
Exemplo: Variável padronizada *X*



Distância entre as observações 1 e 2:

(2 – (–3)) = 5

Exemplo: Variável padronizada 2 * *X*

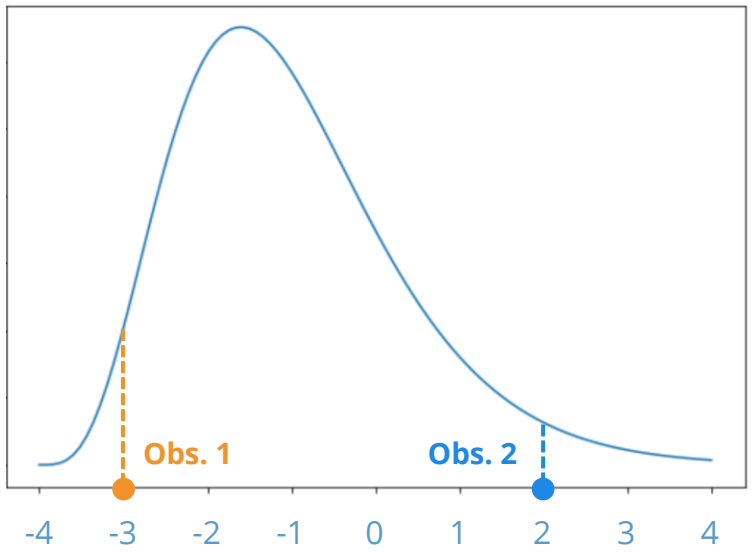


Distância entre as observações 1 e 2:

(4 – (–6)) = 10

Ampliando ou Reduzindo a Escala

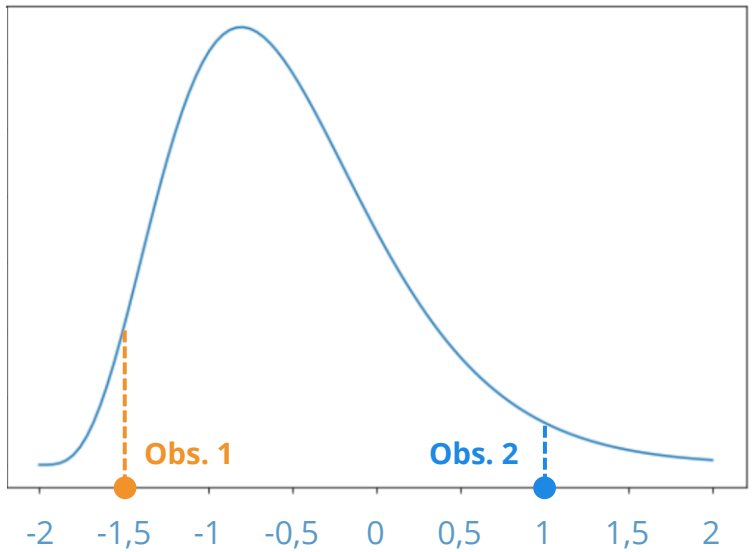
Exemplo: Variável padronizada *X*



Distância entre as observações 1 e 2:

(2 – (–3)) = 5

Exemplo: Variável padronizada 0,5 * *X*



Distância entre as observações 1 e 2:

(1 – (–1,5)) = 2,5

6. Algoritmo Hierárquico

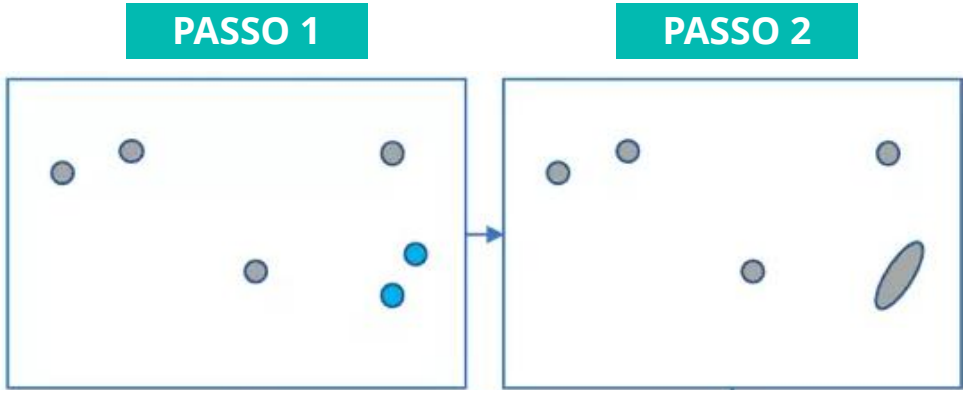
Como Funciona o Algoritmo Hierárquico?

Feita a introdução a respeito das medidas de distância e da padronização de variáveis, já temos o insumo necessário para começar a estudar os algoritmos de **agrupamento**.

O primeiro deles é o **hierárquico**, que pode ser utilizado para variáveis **quantitativas** e, portanto, costuma envolver a distância euclidiana. O método funciona a partir da repetição de 2 passos:

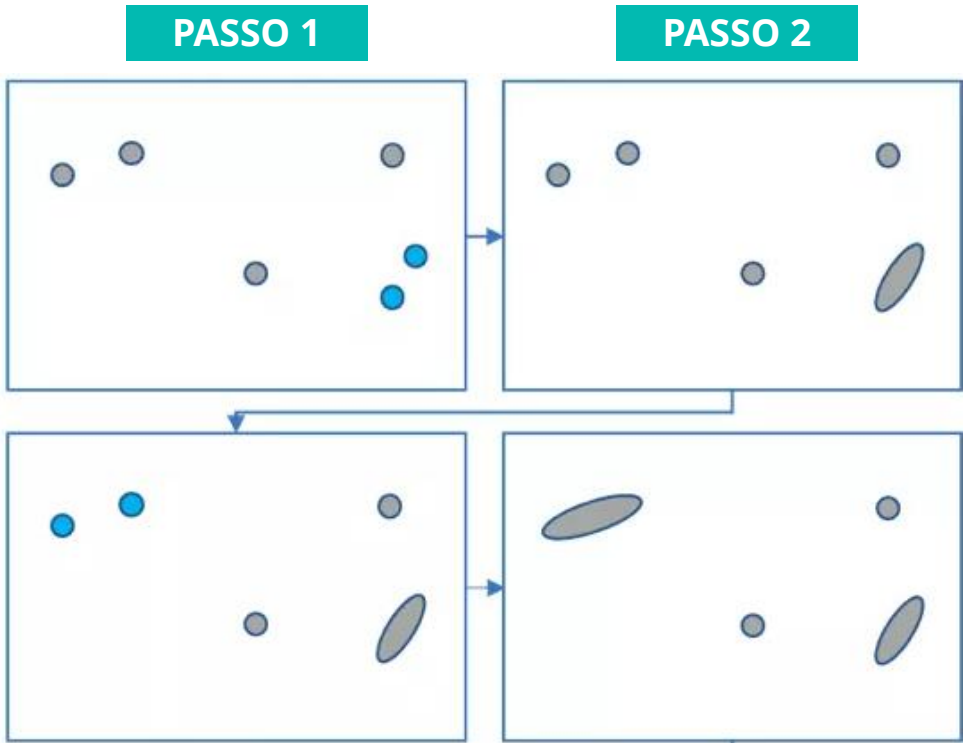
1. Identificar quais são os dois elementos mais próximos entre si, a partir da matriz de distâncias.
2. Unir esses dois elementos em um único. Retornar ao passo 1.

Como Funciona o Algoritmo Hierárquico?



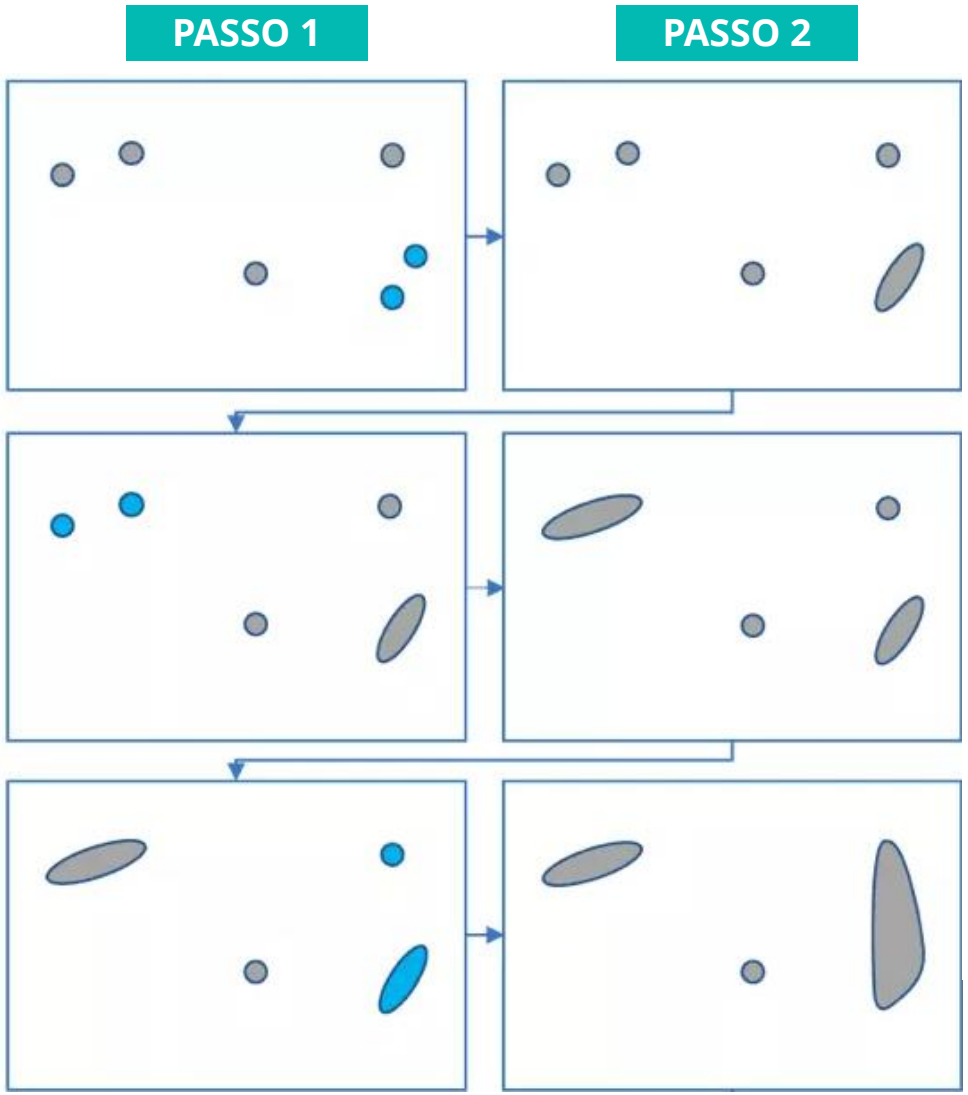
Créditos da imagem:
https://www.displayr.com/what-is-hierarchical-clustering

Como Funciona o Algoritmo Hierárquico?



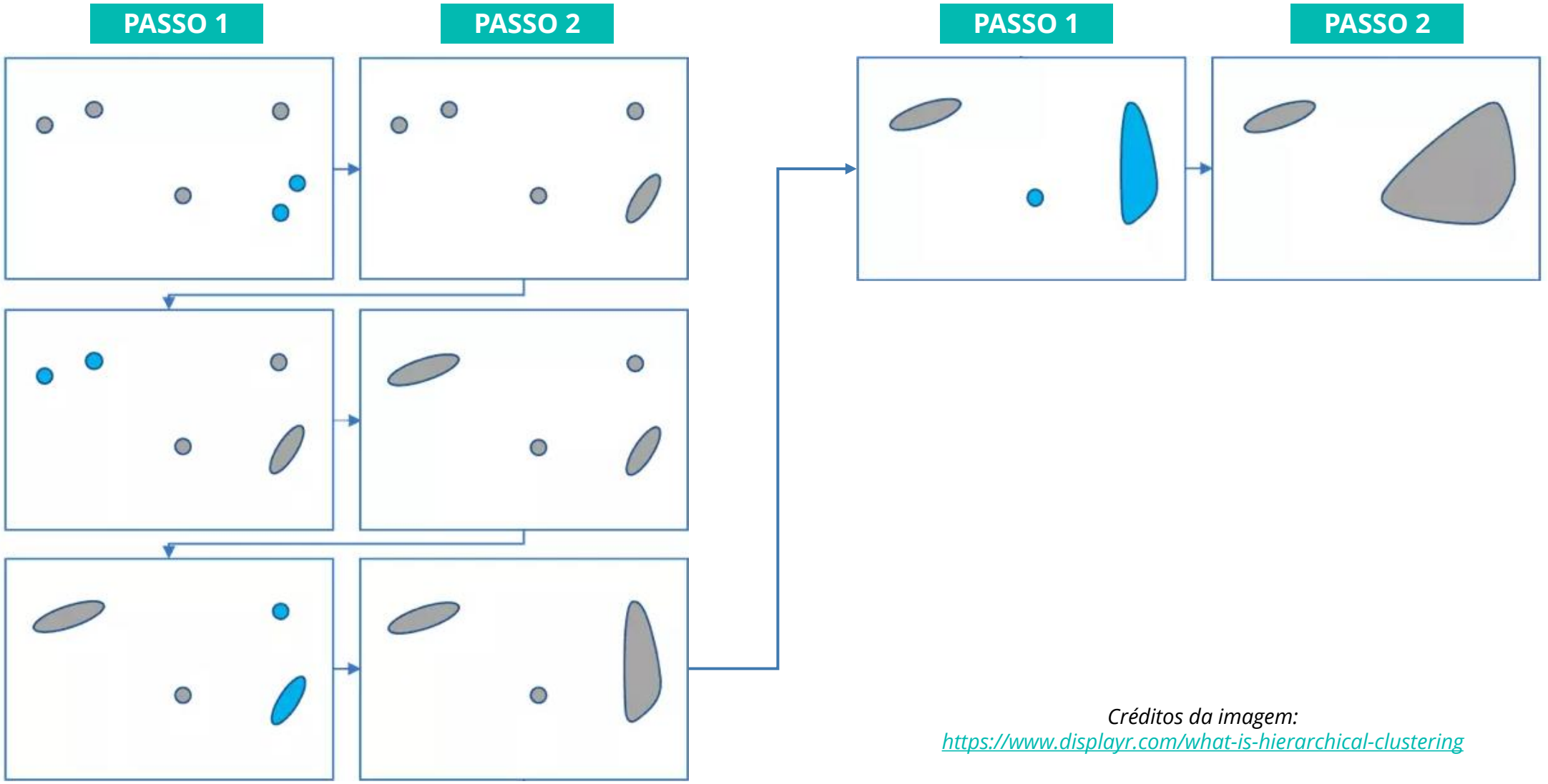
Créditos da imagem:
https://www.displayr.com/what-is-hierarchical-clustering

Como Funciona o Algoritmo Hierárquico?



Créditos da imagem:
https://www.displayr.com/what-is-hierarchical-clustering

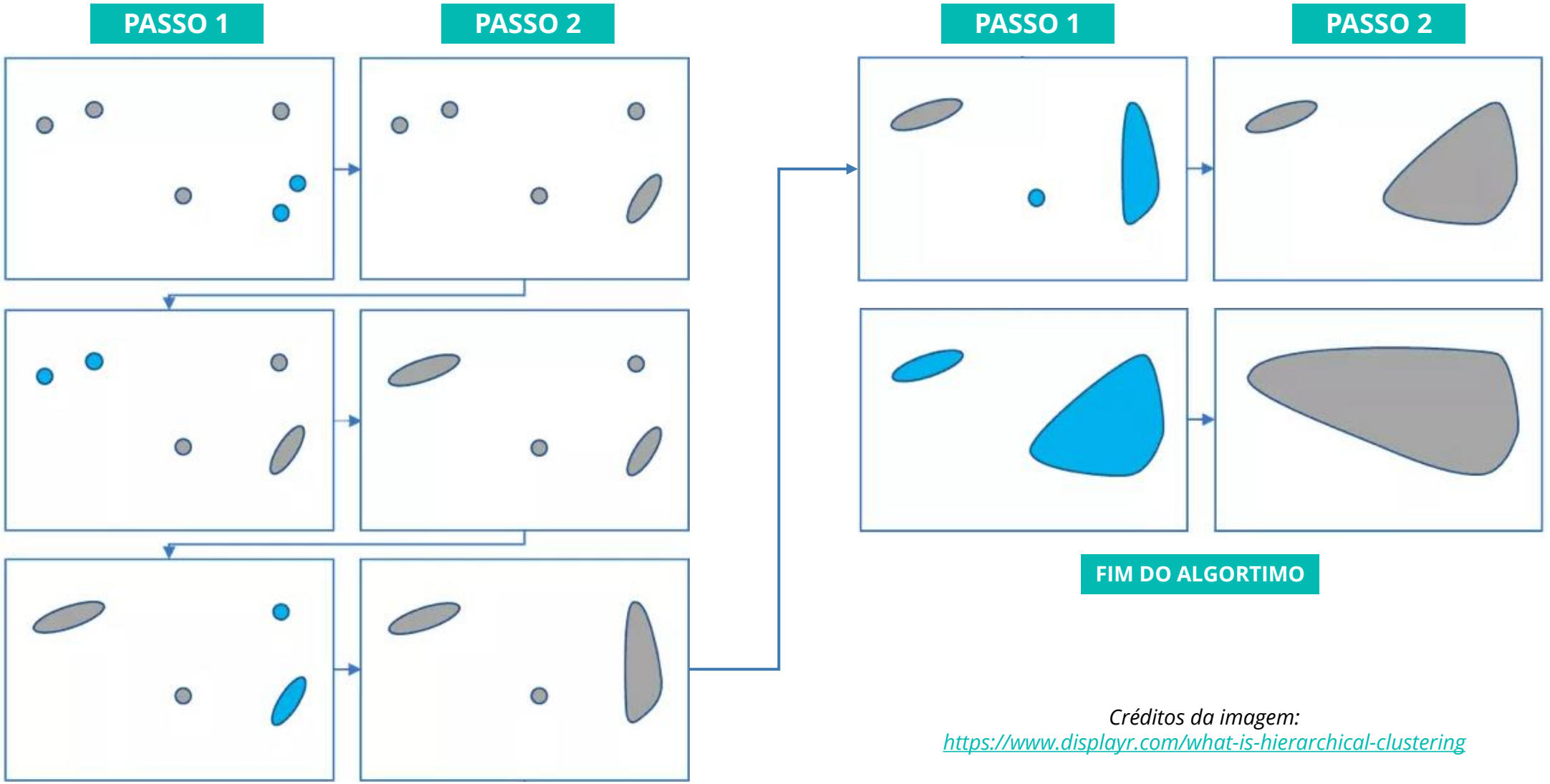
Como Funciona o Algoritmo Hierárquico?



Créditos da imagem:
https://www.displayr.com/what-is-hierarchical-clustering

FIA BUSINESS SCHOOL lab.data

Como Funciona o Algoritmo Hierárquico?



Créditos da imagem:
https://www.displayr.com/what-is-hierarchical-clustering

Como Funciona o Algoritmo Hierárquico?

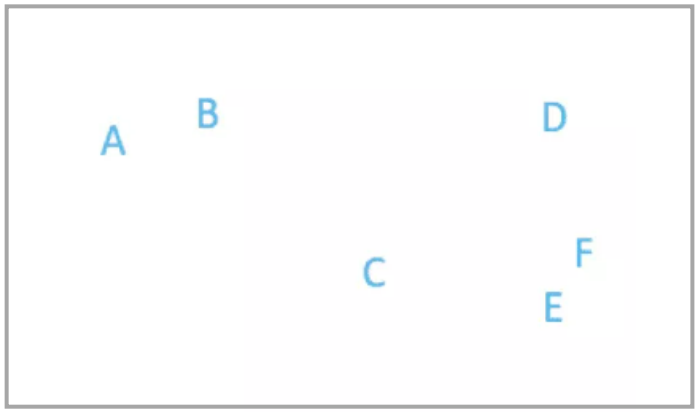
Observe que o método hierárquico se inicia em um cenário com todas as observações individuais, ou seja, em que não foi formado **nenhum *cluster***; e chega em um cenário final em que todas as observações foram agrupadas em um **único *cluster***.

Ao longo desse processo, diversos cenários **intermediários** são identificados, com diferentes quantidades de *clusters*. Cabe a nós identificar qual desses cenários é mais apropriado, usando como ferramenta o gráfico de **dendrograma**.

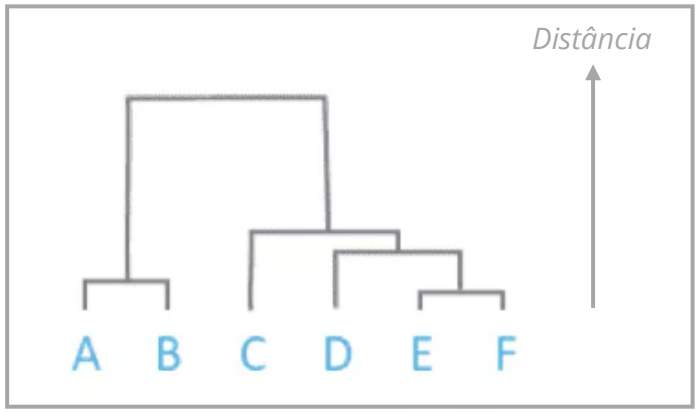
Dendrograma

O **dendrograma** é um gráfico que resume a **sequência de agrupamentos** que foi realizada ao longo do algoritmo hierárquico *versus* o valor da distância entre os elementos agrupados em cada passo.

Observações



Dendrograma

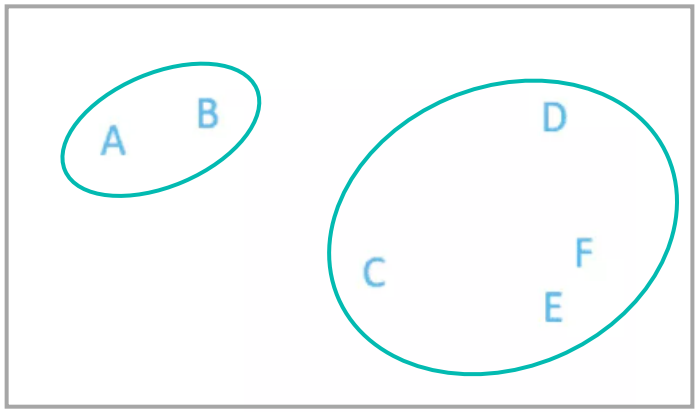


Lendo a sequência de agrupamentos no dendrograma de baixo para cima, buscamos identificar uma região de **estabilidade**. Ou seja, a quantidade ideal de clusters é aquele que antecede o instante em que é necessário **elevar muito** a medida de distância para um novo agrupamento.

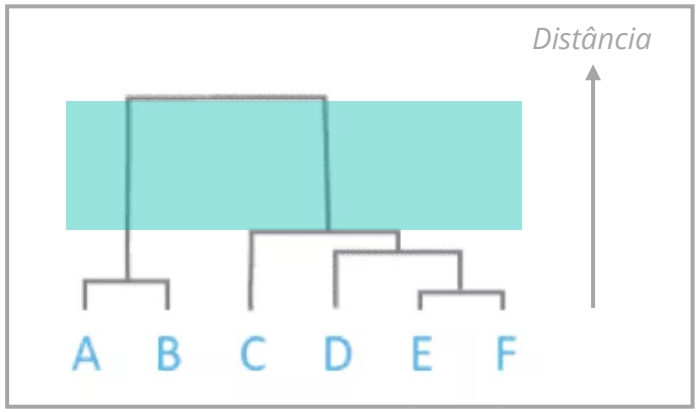
Dendrograma

O **dendrograma** é um gráfico que resume a **sequência de agrupamentos** que foi realizada ao longo do algoritmo hierárquico *versus* o valor da distância entre os elementos agrupados em cada passo.

Observações



Dendrograma



Neste exemplo, o cenário em que o dendrograma atingiu estabilidade foi o de **2 *clusters***.

Este é o cenário **estatisticamente ótimo**, mas outros cenários também podem ser adotados se forem interessantes do ponto de interpretação prática.

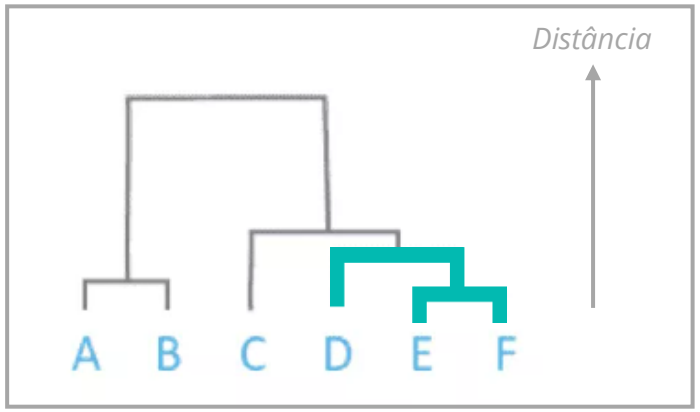
FIA BUSINESS SCHOOL lab.data

Dendrograma

Examinemos as observações **D, E** e **F**. Na matriz de distâncias, temos as distâncias entre os pares **(D, E)**, **(D, F)** e **(E, F)**. Porém, qual é a distância entre **D** e o *cluster* previamente formado com **E** e **F**?

Observações

Dendrograma



A cada agrupamento realizado, o método hierárquico **recalcula** a matriz de distâncias, considerando as duas observações recém-agrupadas como uma única.

Este recálculo é realizado considerando diferentes **critérios de ligação** possíveis.

Critérios de Ligação

Os **critérios de ligação** mais comuns são:

- ***Complete***: assume a **maior** distância entre os dois elementos agrupados e cada um dos demais.

  Exemplo:

dist (D, **E**) = 3	Ao agrupar **E** e **F**, a nova distância entre D e o grupo **(E, F)** é:
dist (D, **F**) = 2	*máx* (2, 3) = 3

- ***Single***: assume a **menor** distância entre os dois elementos agrupados e cada um dos demais.

  Exemplo:

dist (D, **E**) = 3	Ao agrupar **E** e **F**, a nova distância entre D e o grupo **(E, F)** é:
dist (D, **F**) = 2	*mín* (2, 3) = 2

Critérios de Ligação

Os **critérios de ligação** mais comuns são:

- ***Average***: assume a **média** entre as distâncias dos dois elementos agrupados e os demais.

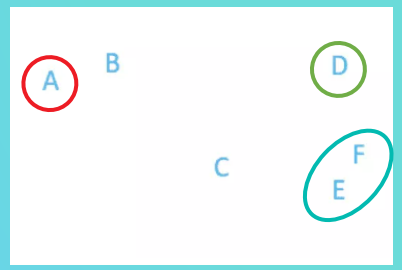
  Exemplo:

  dist (D, **E**) = 3
  dist (D, **F**) = 2

  Ao agrupar **E** e **F**, a nova distância entre D e o grupo **(E, F)** é:
  média (2, 3) = 2,5

- ***Ward***: assume uma métrica de quanto aumentará a **heterogeneidade** (variância) dentro dos *clusters*, caso seja feita uma junção entre os dois elementos agrupados e cada um dos demais.

  Exemplo:



Ao agrupar **E** e **F**, a nova distância entre D e o grupo **(E, F)** será menor do que a distância entre A e o grupo **(E, F)**, pois o futuro grupo (D, E, F) será mais homogêneo do que o grupo (A, E, F).

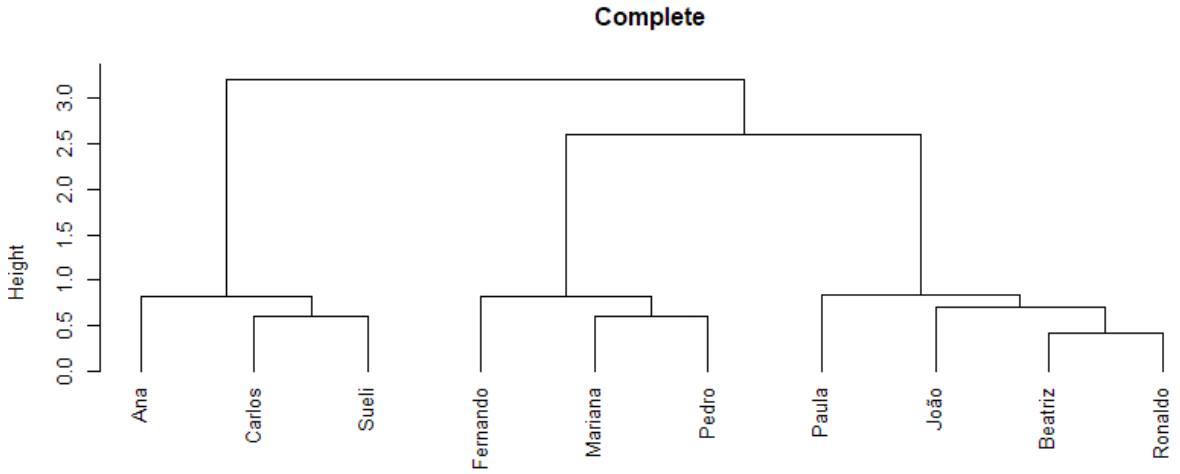
Critérios de Ligação

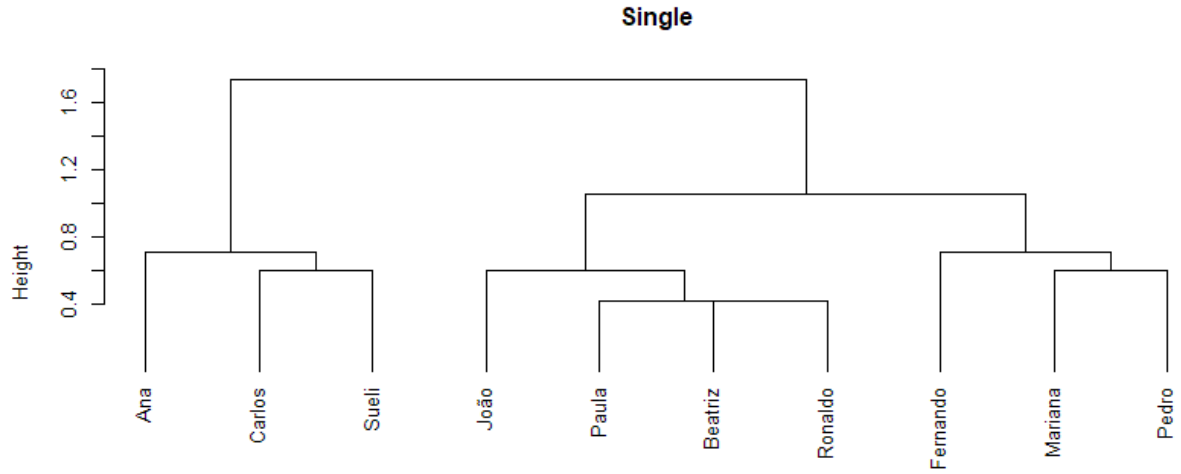
Dessa forma, podemos reescrever os **passos** do algoritmo hierárquico de forma um pouco mais detalhada:

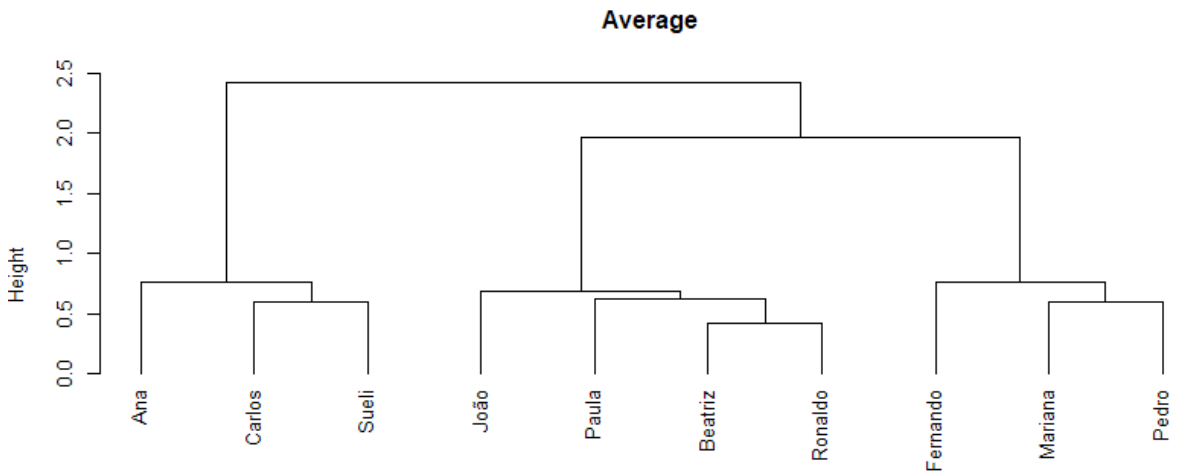
1. Identificar quais são os dois elementos mais próximos entre si, a partir da matriz de distâncias.
2. Unir esses dois elementos em um único **e recalcular a matriz de distâncias, considerando o critério de ligação adotado.** Retornar ao passo 1.

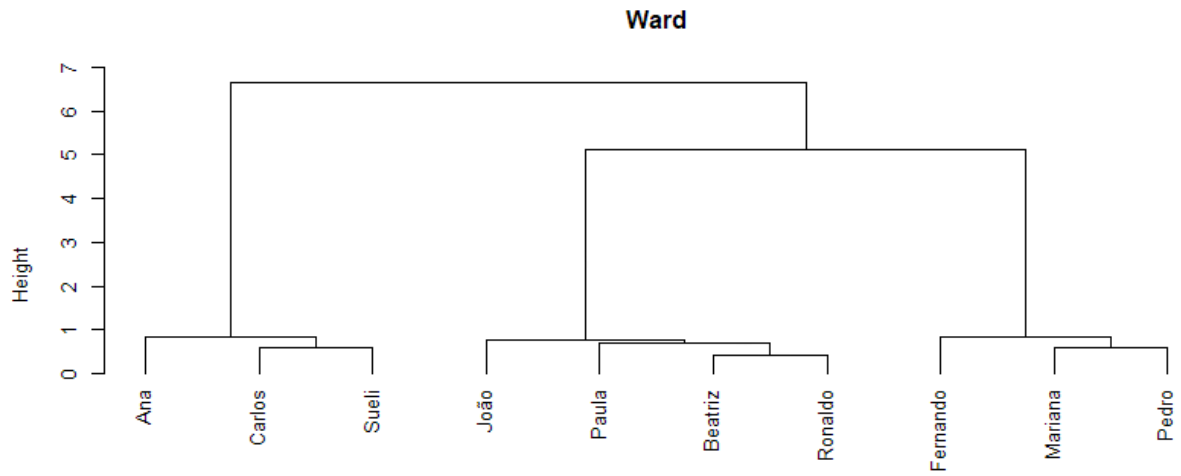
Case: Avaliação de Candidatos

Executando o algoritmo hierárquico para o *case* de avaliação de candidatos, com diferentes **critérios de ligação**:



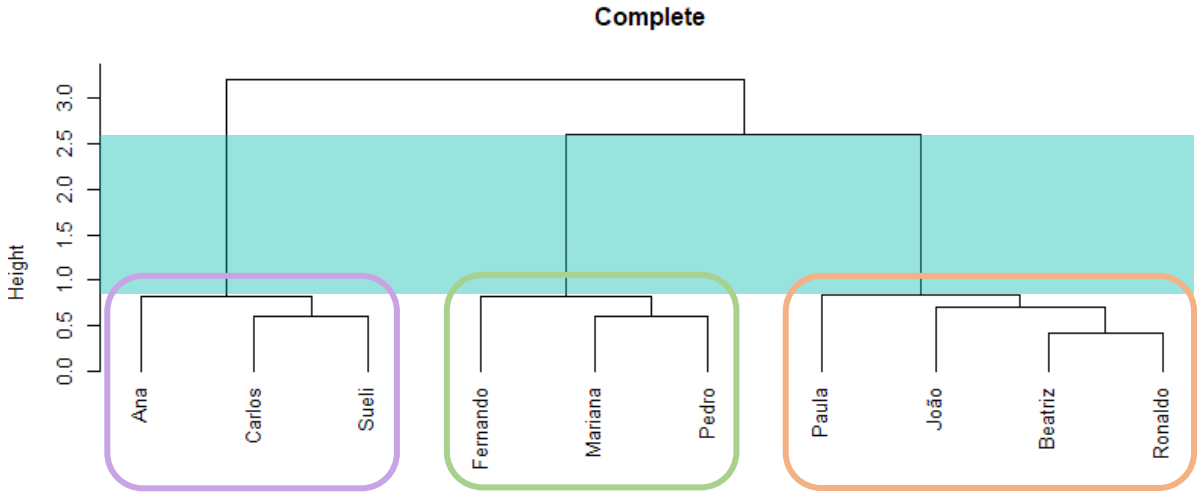


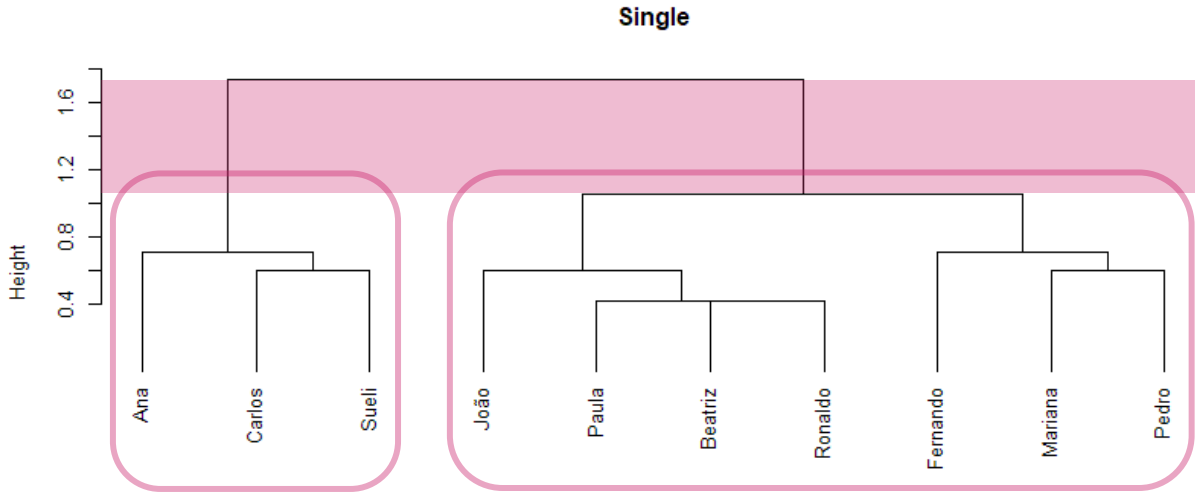


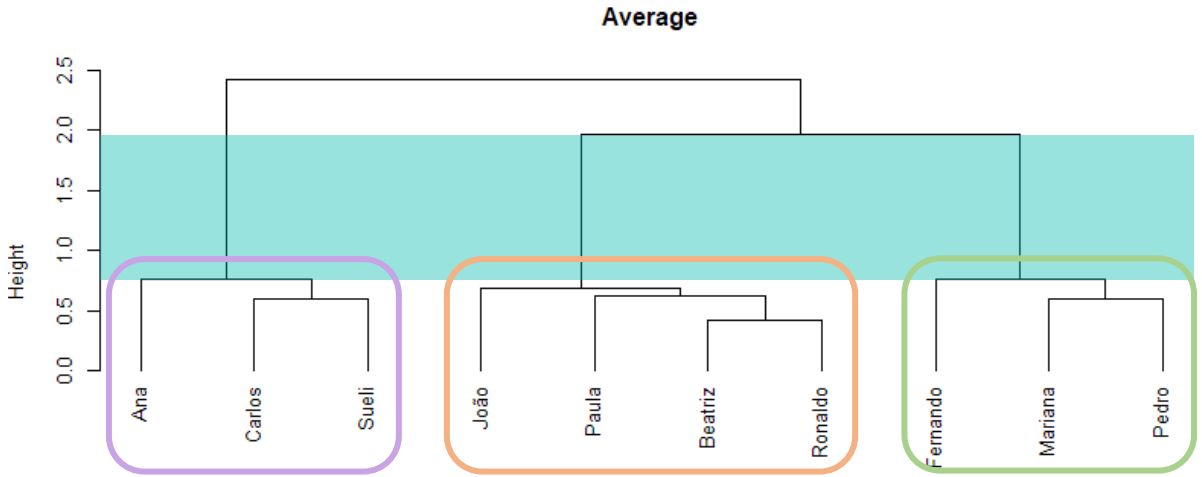


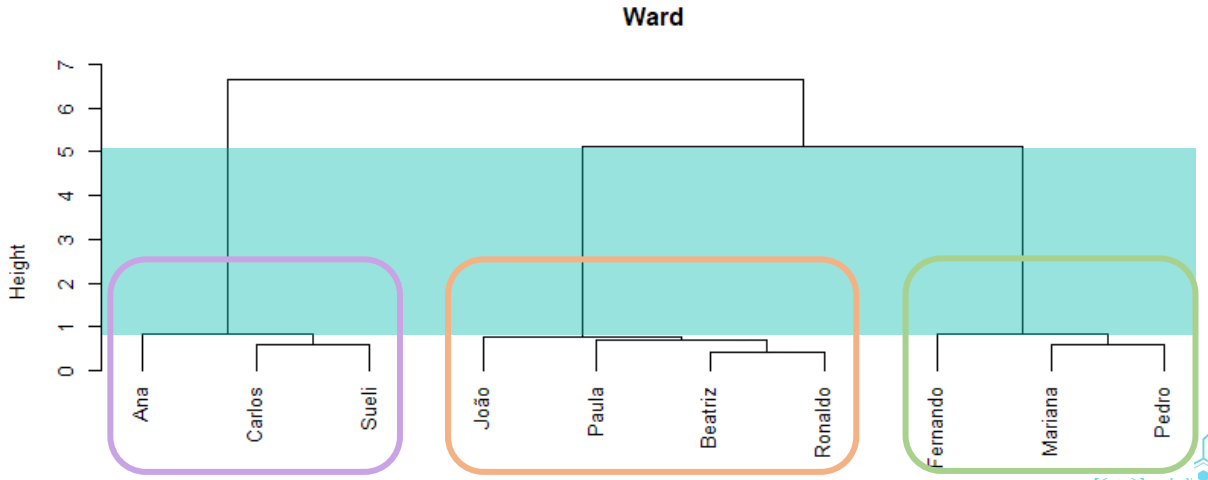
Case: Avaliação de Candidatos

A maioria dos métodos identifica os mesmos **3 *clusters*** que havíamos pré-identificado a partir da análise visual.









FIA BUSINESS SCHOOL lab.data

Case: Avaliação de Candidatos

A última etapa consiste em caracterizar os grupos encontrados a partir de uma **análise exploratória**.

Segmentação hierárquica com *k* = 3 clusters *(ligações complete, average ou Ward)*

Cluster	Tempo médio de experiência	Qtde. média de cursos/especializações	Idade média	Salário médio atual	Senioridade
Cluster 1 (Ana, Carlos, Sueli)	10	8	35	R$ 8.600	100% de sêniores
Cluster 2 (Fernando, Mariana, Pedro)	9	1	32	R$ 6.733	33% de sêniores
Cluster 3 (Beatriz, João, Paula, Ronaldo)	3	4	25	R$ 6.250	25% de sêniores

Se o recrutador tiver que escolher **apenas 1 *cluster*** de candidatos para prosseguir no processo seletivo, qual ele escolheria? Obviamente, essa decisão dependerá do perfil de profissional que se está buscando, do budget de remuneração para a vaga, entre outros fatores.

Arquivo: Avaliacao_Candidatos (.txt)

FIA BUSINESS SCHOOL lab.data

Resumo do Procedimento

Por fim, sugere-se o seguinte **procedimento** para realização do agrupamento hierárquico com variáveis quantitativas, após a etapa de planejamento e definição das variáveis da segmentação:

1. Realizar a **padronização** das variáveis.
2. Calcular a matriz de **distâncias euclidianas**.
3. Aplicar o **algoritmo** (via *software*), testando um ou mais **critérios de ligação**.
4. Avaliar os **dendrogramas** para cada critério de ligação, e identificar um ou mais cenários de maior estabilidade (exemplo: *complete* com 4 *clusters* e *Ward* com 3 *clusters*).
5. Para cada cenário potencial, realizar a **análise exploratória/descritiva** dos *clusters* e interpretá-los. Lembrando que não há uma única solução correta; diferentes agrupamentos podem ser bons a depender da interpretabilidade dos *clusters*!

7. Algoritmos de Partição

Como Funcionam os Algoritmos de Partição?

Uma outra classe de algoritmos voltados para segmentação de observações é a de **partição**.

A principal diferença inicial entre o algoritmo hierárquico e os algoritmos de partição é que estes requerem a definição prévia de **quantos *clusters*** deseja-se identificar, geralmente denotada pela letra ***k***.

O passo-a-passo geral dos algoritmos de partição é o seguinte:

1. Define-se *k* pontos de referência iniciais e aleatórios, para cada um dos *k clusters*.
2. Todas as observações são atribuídas ao *cluster* mais próximo, de acordo com alguma métrica de distância.
3. Os *k* pontos de referência iniciais são atualizados, de acordo com a distribuição das observações que foram atribuídas a cada *cluster*, e retorna-se para o passo 2.

Os passos 2 e 3 são **repetidos consecutivamente** até que o algoritmo atinja um cenário de estabilidade, ou seja, que mais nenhuma observação mude de *cluster*.

Como Funcionam os Algoritmos de Partição?

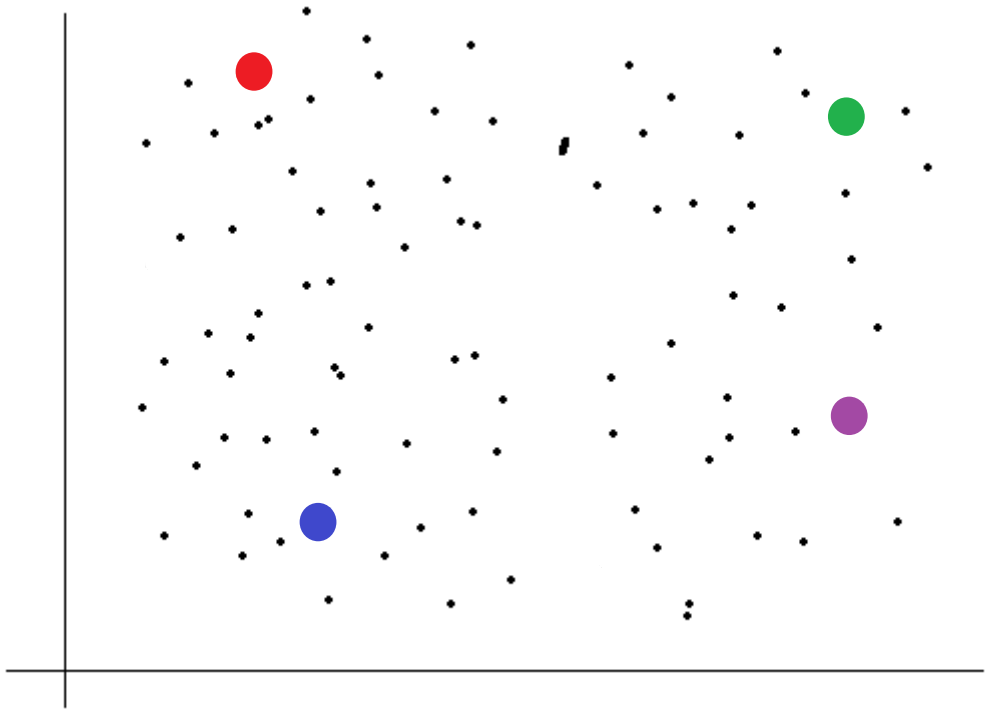
Os algoritmos de partição mais comuns são:

- ***K*-médias**: Apropriado para variáveis quantitativas; utiliza a distância euclidiana.
- ***K*-medoides**: Apropriado para quaisquer tipos de variáveis; utiliza, em geral, a distância de Gower.

Algoritmo K-Médias

Passo 1: Define-se *k* pontos de referência iniciais e aleatórios no espaço, para cada um dos *k clusters*.

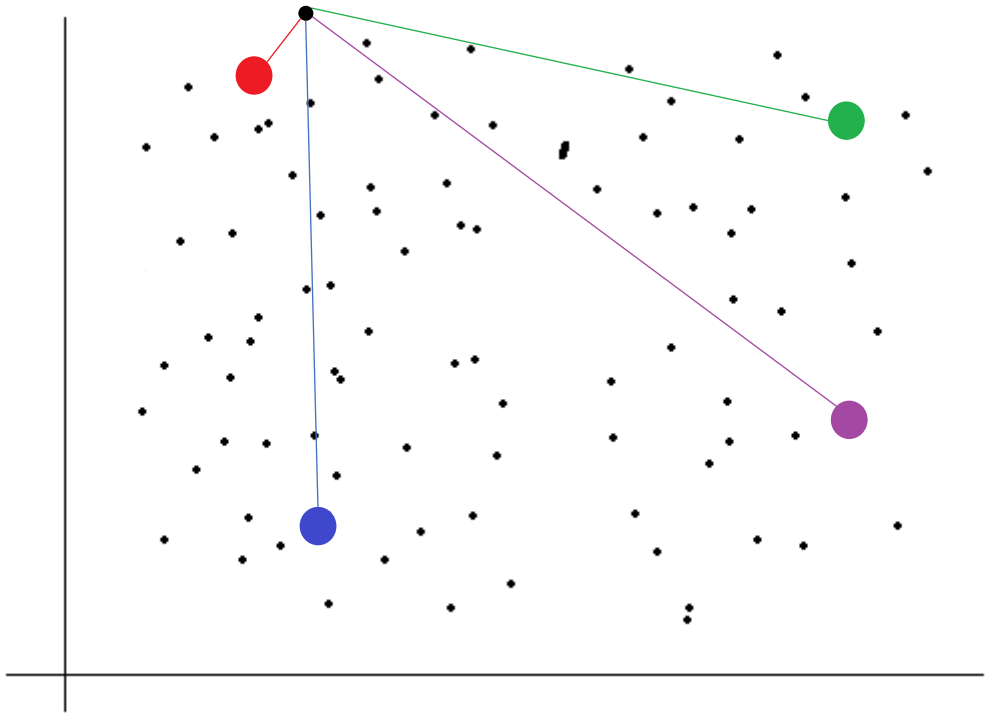
No exemplo abaixo, nosso espaço é de **2 dimensões** (ou seja, 2 variáveis) e a quantidade de *clusters* é ***k* = 4**.



No *k*-médias, os pontos de referência de cada *cluster* são também chamados de **centroides**.

Passo 2: Todas as observações são atribuídas ao *cluster* mais próximo, de acordo a **distância euclidiana**.

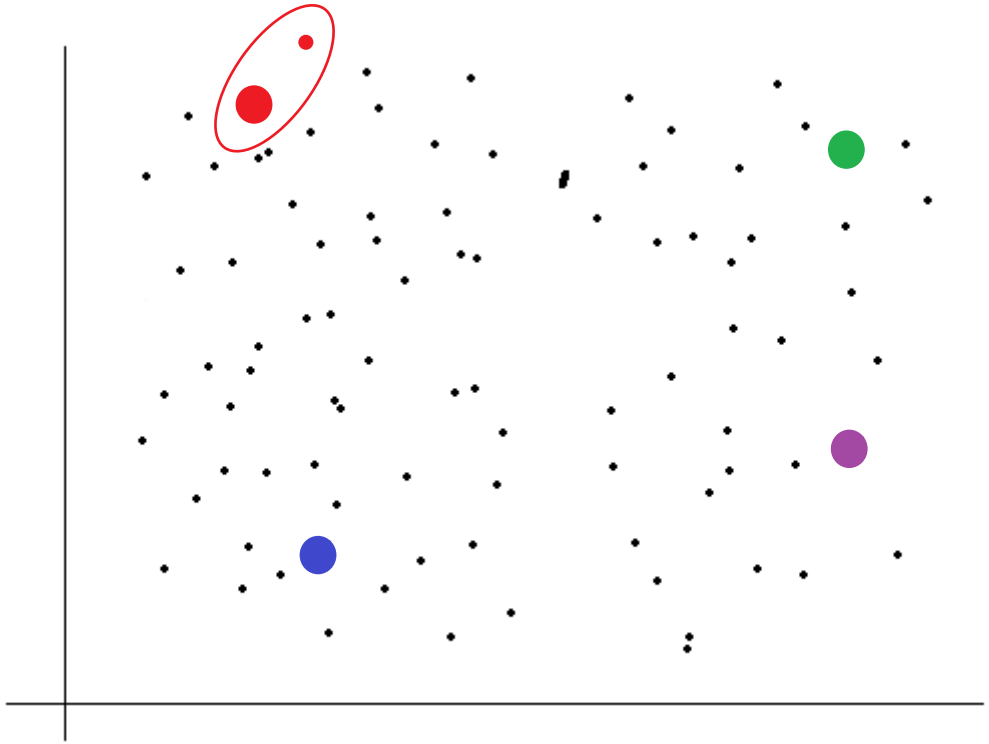
A observação destacada em preto, abaixo, está mais próxima de qual *cluster*?



Algoritmo K-Médias

Passo 2: Todas as observações são atribuídas ao *cluster* mais próximo, de acordo a **distância euclidiana**.

Por estar mais próxima do *cluster* com centroide **vermelho**, a observação passa a pertencer a este *cluster*.

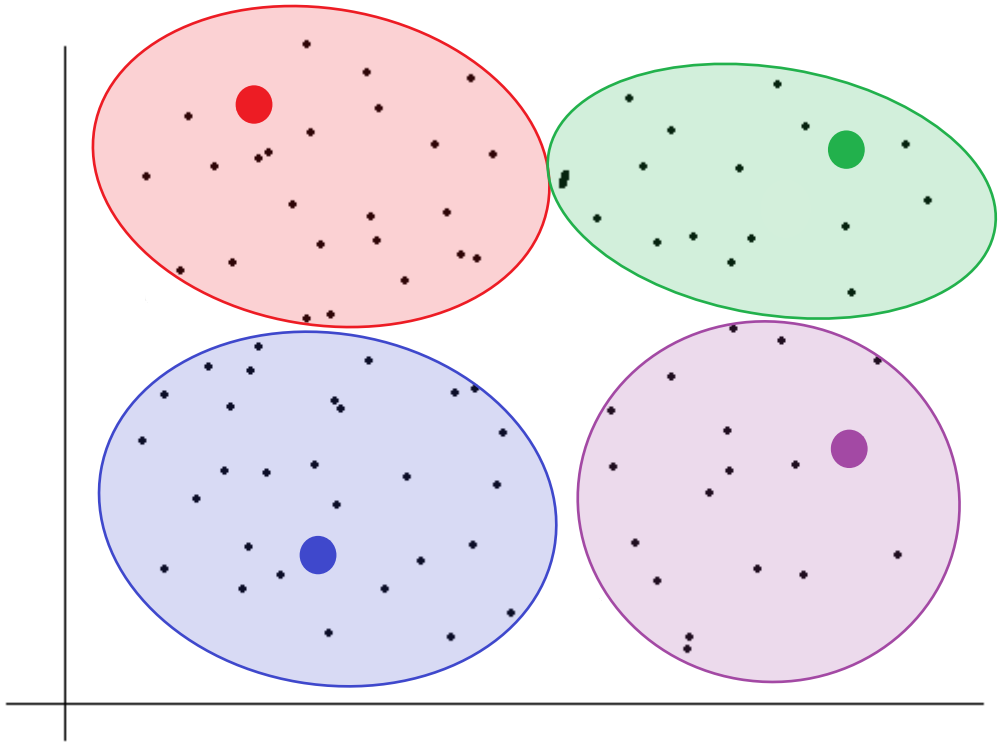


FIA BUSINESS SCHOOL lab.data

Algoritmo K-Médias

Passo 2: Todas as observações são atribuídas ao *cluster* mais próximo, de acordo a **distância euclidiana**.

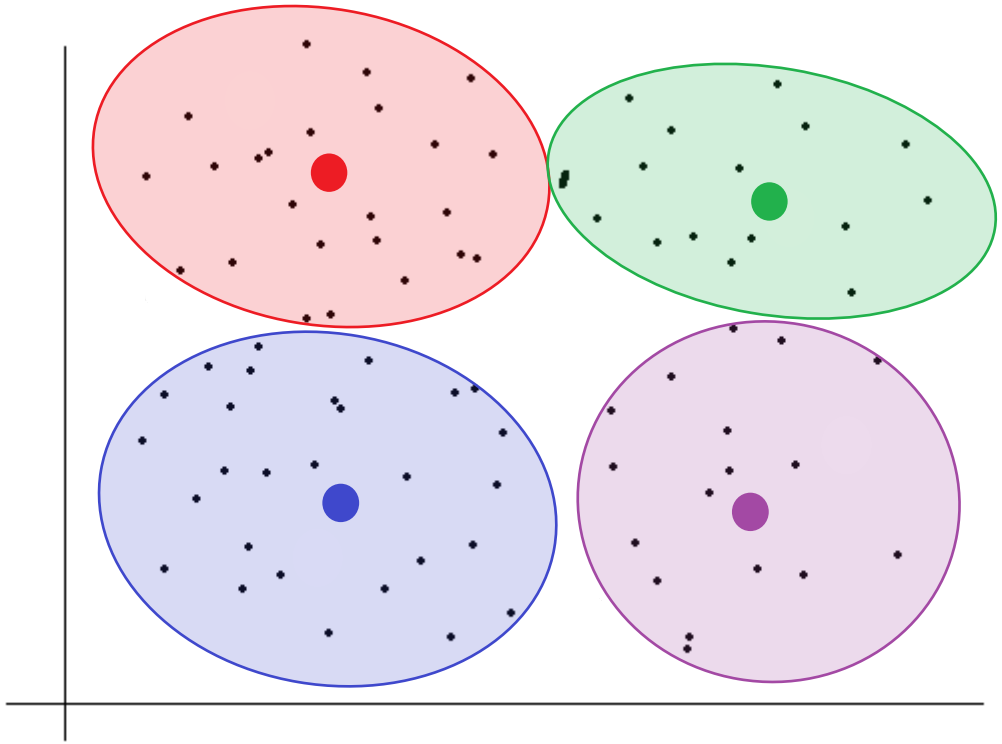
Ao final deste passo, todas as observações terão sido atribuídas a algum *cluster*.



Algoritmo K-Médias

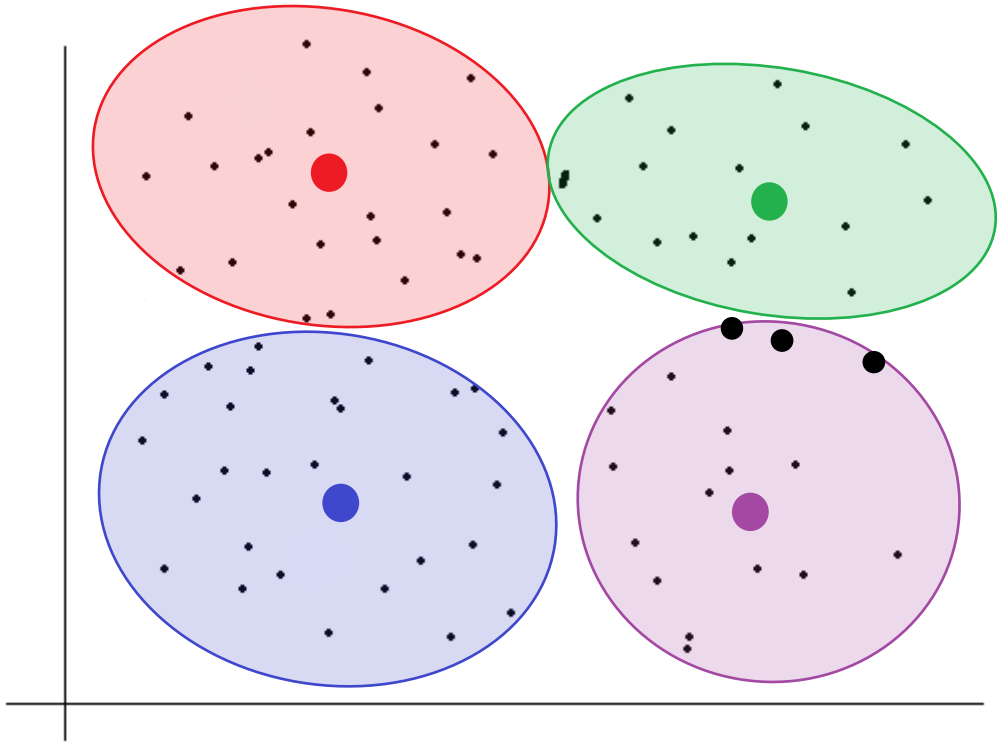
Passo 3: Os *k* centroides são atualizados, a partir do **valor médio** das observações em cada *cluster*.

Este passo atua como uma "correção" dos pontos de referência iniciais, que haviam sido escolhidos aleatoriamente.



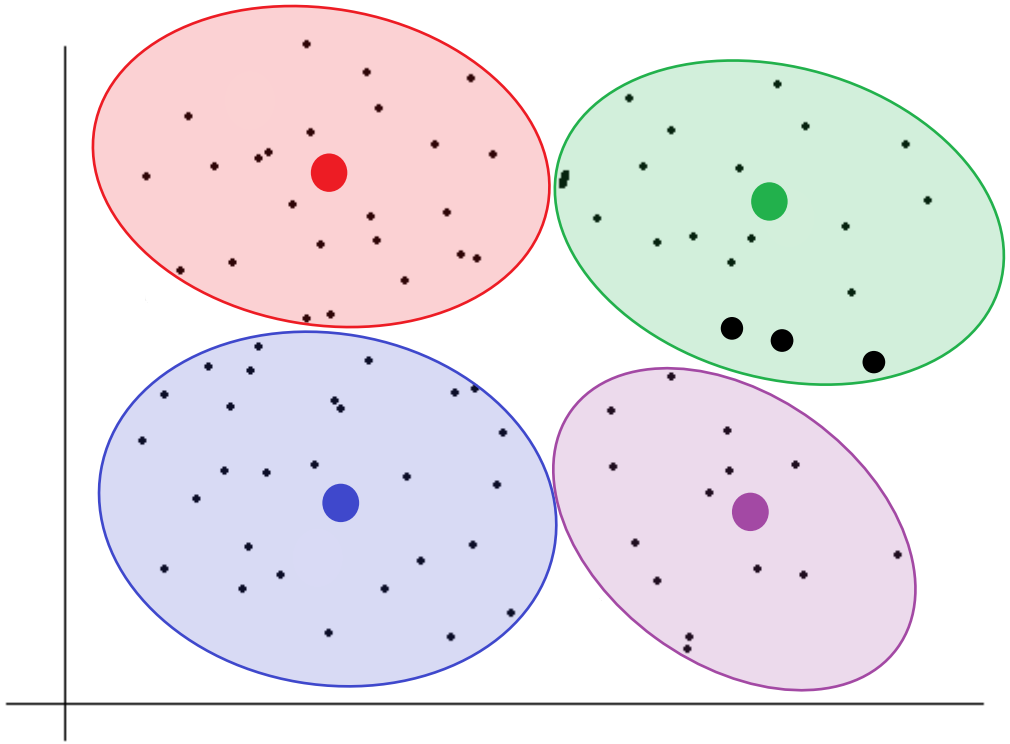
Algoritmo K-Médias

Note que apesar de fazerem parte do *cluster* roxo, as observações destacadas abaixo estão mais próximas, agora, do centroide do grupo verde. Por isso, é necessário repetir os passos 2 e 3 sucessivamente para **otimizar a classificação**, até que nenhuma observação tenha que ser realocada.

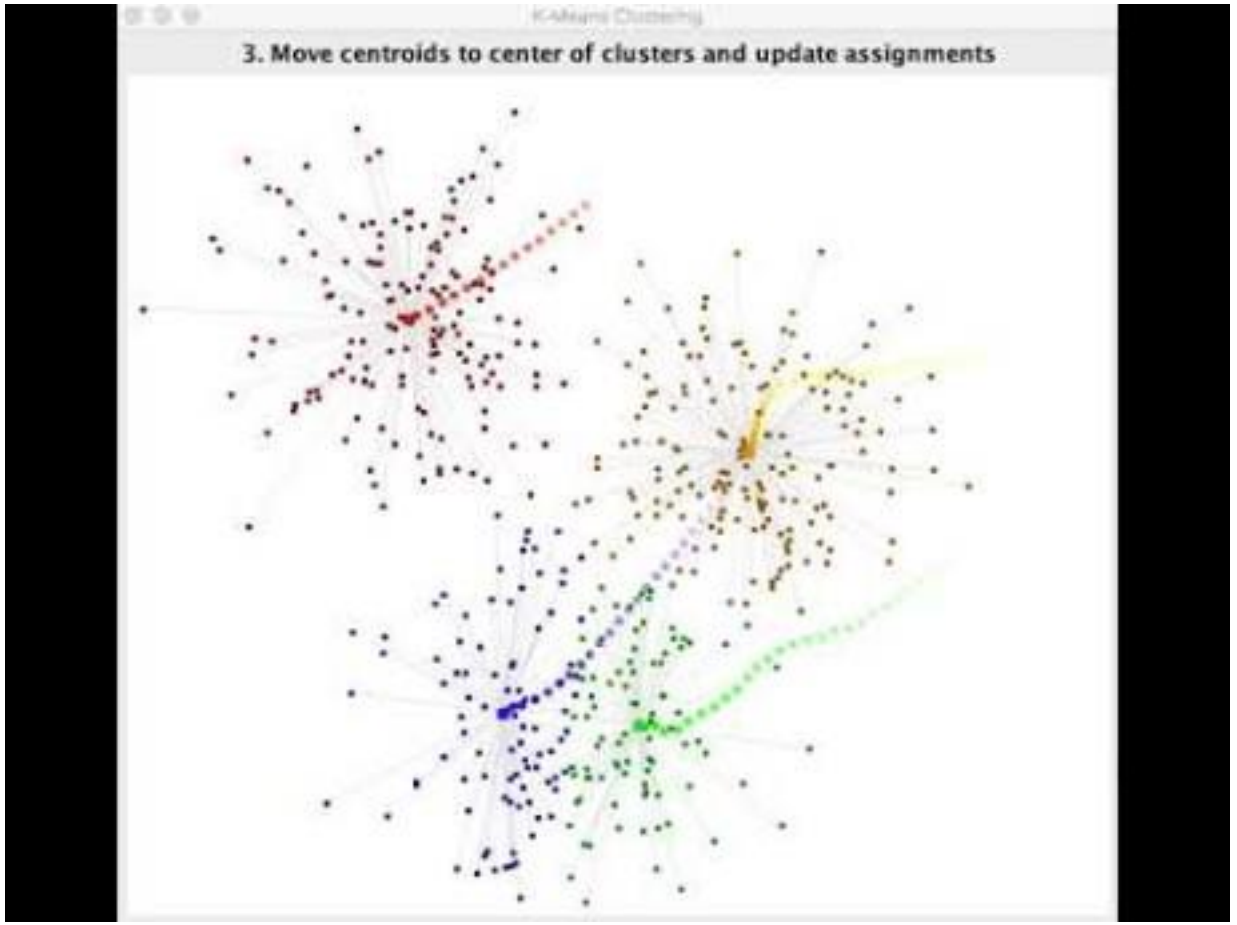


Algoritmo K-Médias

Note que apesar de fazerem parte do *cluster* roxo, as observações destacadas abaixo estão mais próximas, agora, do centroide do grupo verde. Por isso, é necessário repetir os passos 2 e 3 sucessivamente para **otimizar a classificação**, até que nenhuma observação tenha que ser realocada.



Algoritmo K-Médias



Exemplo do processo iterativo descrito anteriormente

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

Algoritmo K-Medoides

O algoritmo ***k*-medoides** funciona de forma análoga ao *k*-médias, mas permite misturar variáveis quantitativas e variáveis qualitativas. Suas nuances são apresentadas a seguir.

1. *Definição de k pontos de referência iniciais e aleatórios*

Em vez de serem sorteados *k* valores numéricos, são sorteadas *k* **observações reais** da base de dados. Aqui, os pontos de referência são denominados **medoides**.

2. *Atribuição das observações ao cluster mais próximo*

Utiliza-se a medida de distância **mais apropriada** para os tipos de variáveis envolvidos (euclidiana, *simple matching*, Gower etc.)

3. *Atualização dos k pontos de referência*

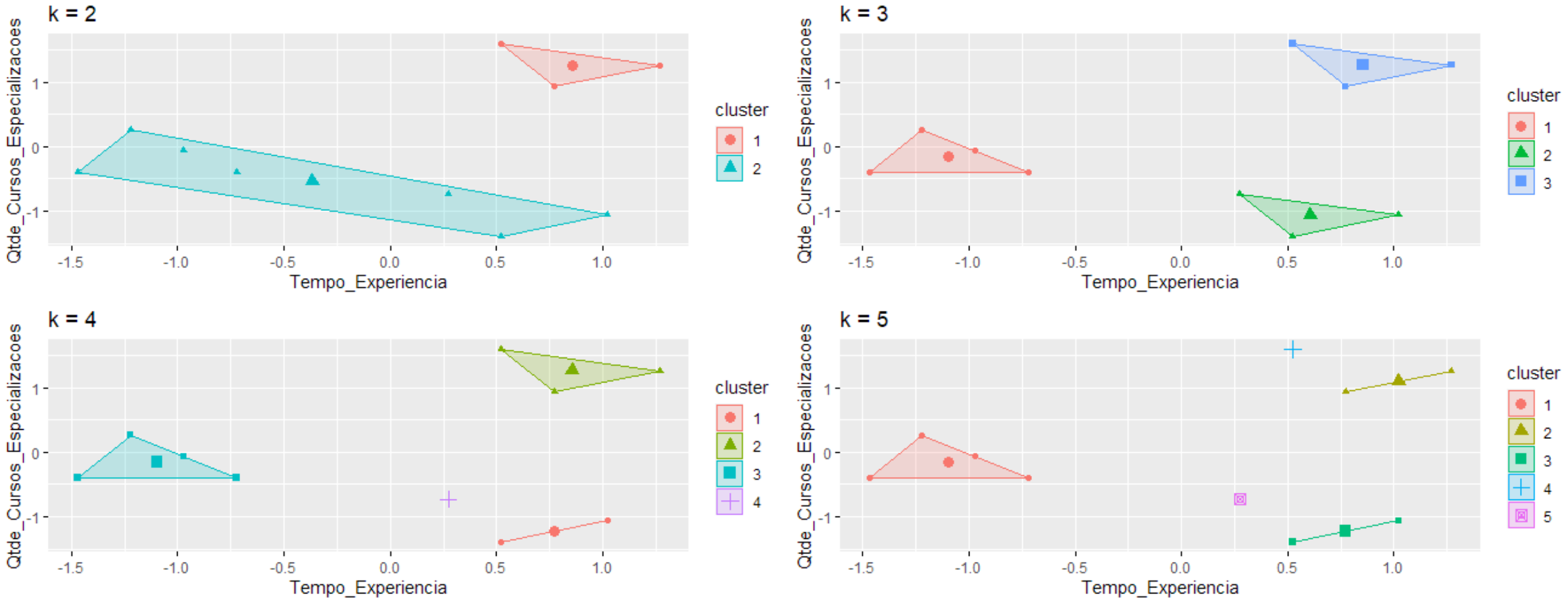
Em vez de os pontos de referência serem as médias das variáveis, aqui, são atualizados de forma a minimizar a **distância média** dos pontos ao medoide, dentro de cada cluster.

Considerações sobre os Métodos de Partição

- O algoritmo ***k*-médias** possui execução computacional **muito mais rápida** do que o algoritmo hierárquico, devido ao fato de não realizar comparações de distância entre todos os pares de observações. Por isso, pode ser utilizados em grandes bases de dados, diferentemente do hierárquico, cuja execução não é viável para volumes maiores que dezenas de milhares de observações.
- Já o algoritmo ***k*-medoides**, devido à sua complexidade no passo 3, requer a matriz de distâncias completa, tal como o algoritmo hierárquico. Portanto, **não é performático** para grandes bases de dados. Para utilizá-lo, recomenda-se trabalhar com uma amostra aleatória das observações.
- Por outro lado, a desvantagem dos métodos de partição consiste em ter que estabelecer previamente a **quantidade de *clusters***.

Case: Avaliação de Candidatos

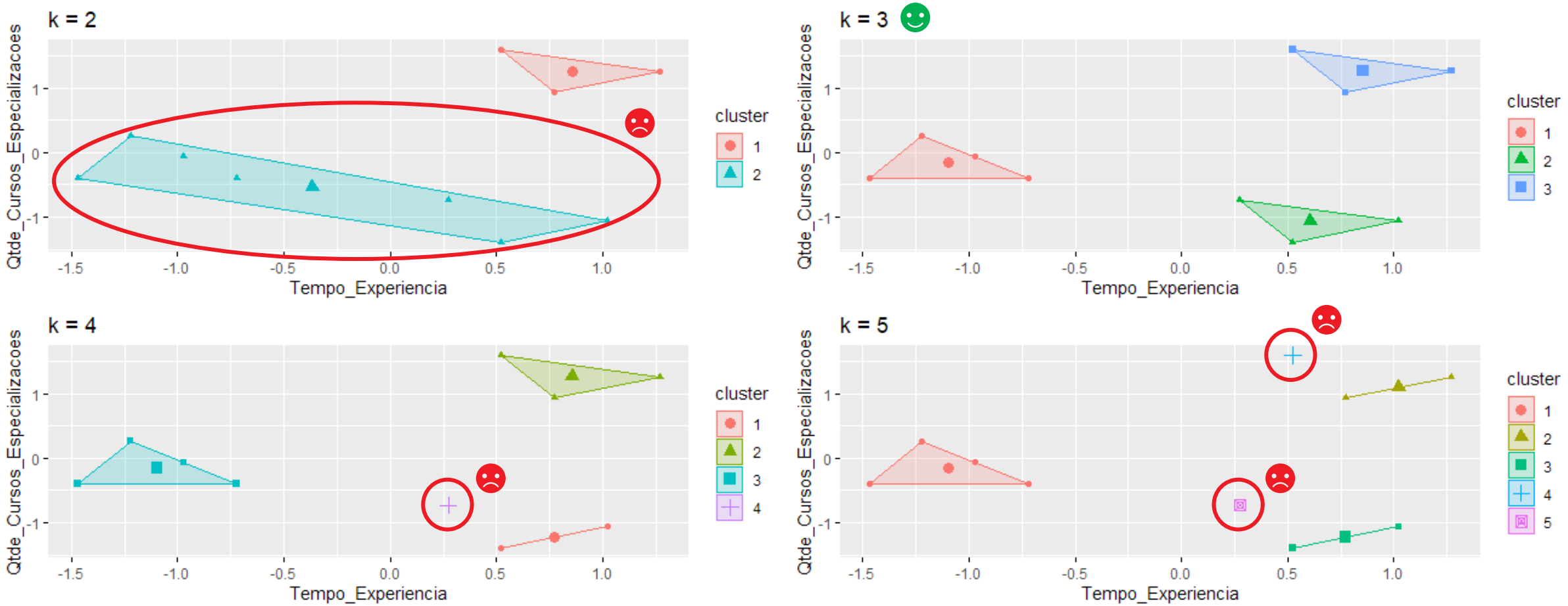
Execução do algoritmo ***k*-médias** para o *case* de avaliação de candidatos, com distância **euclidiana**, considerando as variáveis **quantitativas** (tempo de experiência e quantidade de cursos/especializações) e para *k* = 2, 3, 4 e 5.



FIA BUSINESS SCHOOL lab.data

Case: Avaliação de Candidatos

Visualmente, o cenário com ***k* = 3 *clusters*** parece o ideal, por não agrupar candidatos muito heterogêneos (como para *k* = 2) ou separar candidatos parecidos (como para *k* = 4 ou *k* = 5).



FIA BUSINESS SCHOOL lab.data

Case: Avaliação de Candidatos

Execução do algoritmo ***k*-medoides** para o *case* de avaliação de candidatos, com distância de **Gower**, considerando a junção das 4 variáveis **quantitativas** e **qualitativas**, e para $k = 2, 3$ e 4.

Candidato (a)	Tempo de experiência	Qtde. de cursos/especializações	Área de formação	Nível hierárquico atual	***k* = 2**	***k* = 3**	***k* = 4**
Ana	9	9	Engenharia	Sênior	1	1	1
Beatriz	3	4	Ciência da Computação	Pleno	2	2	2
Carlos	10	7	Análise de Sistemas	Sênior	1	3	3
Fernando	8	2	Ciência da Computação	Sênior	2	3	3
João	1	3	Engenharia	Pleno	2	2	2
Mariana	11	1	Ciência da Computação	Pleno	2	2	2
Paula	4	3	Análise de Sistemas	Sênior	1	3	3
Pedro	9	0	Análise de Sistemas	Pleno	1	3	4
Ronaldo	2	5	Ciência da Computação	Pleno	2	2	2
Sueli	12	8	Engenharia	Sênior	1	1	1

Arquivo: Avaliacao_Candidatos (.txt)

Case: Avaliação de Candidatos

Examinando mais a fundo o cenário com ***k* = 3 *clusters***, que mostrou-se mais apropriado na execução anterior do método *k*-médias.

Candidato (a)	Tempo de experiência	Qtde. de cursos/especializações	Área de formação	Nível hierárquico atual	***k* = 2**	***k* = 3**	***k* = 4**
Ana	9	9	Engenharia	Sênior	1	1	1
Beatriz	3	4	Ciência da Computação	Pleno	2	2	2
Carlos	10	7	Análise de Sistemas	Sênior	1	3	3
Fernando	8	2	Ciência da Computação	Sênior	2	3	3
João	1	3	Engenharia	Pleno	2	2	2
Mariana	11	1	Ciência da Computação	Pleno	2	2	2
Paula	4	3	Análise de Sistemas	Sênior	1	3	3
Pedro	9	0	Análise de Sistemas	Pleno	1	3	4
Ronaldo	2	5	Ciência da Computação	Pleno	2	2	2
Sueli	12	8	Engenharia	Sênior	1	1	1

Cluster 1: Engenheiras sêniores com alto tempo de experiência e alta quantidade de cursos/especializações.
Cluster 2: Profissionais de nível pleno, geralmente com pouca experiência.
Cluster 3: Predominantemente analistas de sistemas sêniores, com bastante tempo de experiência.

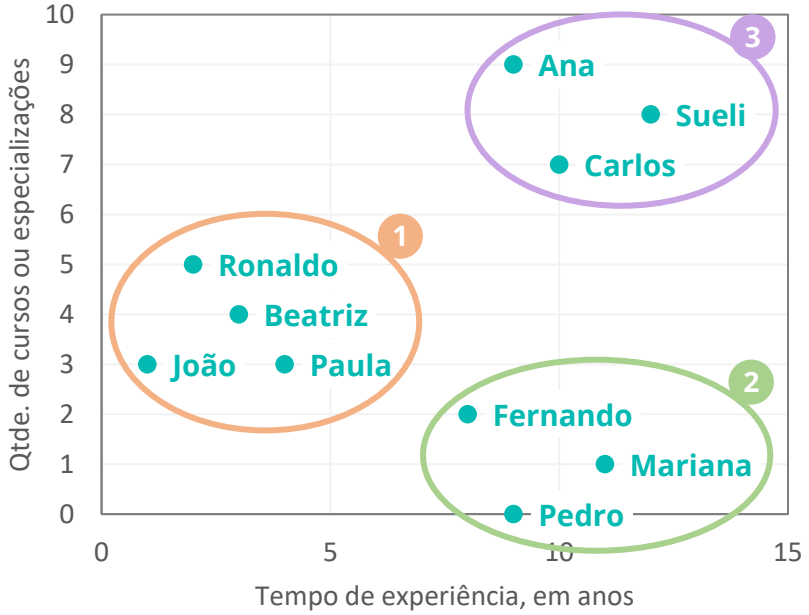
Arquivo: Avaliacao_Candidatos (.txt)

Case: Avaliação de Candidatos

Comparação dos resultados dos dois métodos, para ***k* = 3**, *versus* variáveis quantitativas.

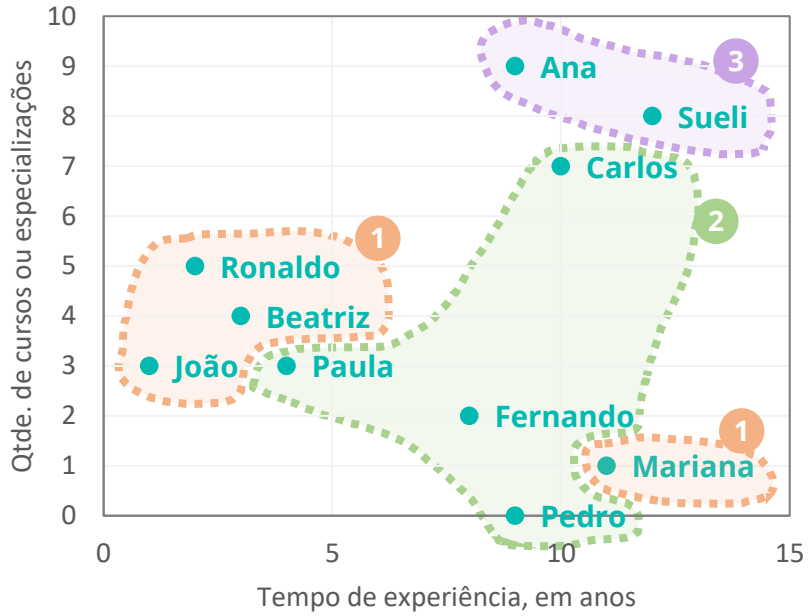
Algoritmos hierárquico e *k*-médias

***Variáveis**: Tempo de experiência e qtde. de cursos*



Algoritmo *k*-medoides

***Variáveis**: Tempo de experiência, qtde. de cursos, área de formação e nível hierárquico*



Note que o *k*-medoides juntou candidatos que não são tão semelhantes entre si no que diz respeito às variáveis **quantitativas** (ex.: Mariana), porque incorporamos na segmentação, também, as variáveis **qualitativas**.

8. *Cases* Adicionais

Case: Hábitos Alimentares

Os dados abaixo são provenientes de uma pesquisa de consumo de alimentos em 25 países da Europa ao longo de determinado período. Ao todo, o consumo de nove grupos de alimentos foi analisado. Temos como objetivo agrupar os países que possuem comportamentos de alimentação semelhantes.

Adaptado a partir de: DASL *(The Data and Story Library)*

Variável	Descrição
pais	Nome do país
carne_vermelha	Índice médio de consumo anual de **carne vermelha**, per capita (em quilos)
carne_branca	Índice médio de consumo anual de **carne branca**, per capita (em quilos)
ovos	Índice médio de consumo anual de **ovos**, per capita (em unidades)
leite	Índice médio de consumo anual de **leite**, per capita (em litros)
peixes	Índice médio de consumo anual de **peixes**, per capita (em quilos)
cereais	Índice médio de consumo anual de **cereais**, per capita (em quilos)
carboidratos	Índice médio de consumo anual de **carboidratos**, per capita (em quilos)
graos	Índice médio de consumo anual de **grãos**, per capita (em quilos)
frutas_legumes	Índice médio de consumo anual de **frutas e legumes**, per capita (em quilos)

Arquivo: Habitos_Alimentares (.txt)

Case: Hábitos Alimentares

Os dados abaixo são provenientes de uma pesquisa de consumo de alimentos em 25 países da Europa ao longo de determinado período. Ao todo, o consumo de nove grupos de alimentos foi analisado. Temos como objetivo agrupar os países que possuem comportamentos de alimentação semelhantes.

Adaptado a partir de: DASL *(The Data and Story Library)*

pais	carne_vermelha	carne_branca	ovos	leite	peixes	cereais	carboidratos	graos	frutas_legumes
Albania	86,0	14,2	34,9	173,5	1,9	81,7	15,8	97,8	31,4
Alemanha	97,1	127,1	286,6	366,5	32,3	35,9	137,2	26,7	70,2
Austria	75,8	142,4	300,5	388,0	20,0	54,1	95,0	23,1	79,5
Belgica	115,0	94,6	286,6	341,2	42,8	51,4	150,4	37,3	73,9
Bulgaria	66,4	61,0	111,8	161,8	11,4	109,5	29,0	65,8	77,6
Croacia	37,5	50,8	83,9	185,2	5,7	108,0	79,2	101,4	59,1
Dinamarca	90,3	109,8	258,6	487,4	94,2	42,3	126,7	12,4	44,4
Eslovaquia	82,6	115,9	195,7	243,7	19,0	66,2	132,0	19,6	73,9
Espanha	60,5	34,6	216,7	167,7	66,6	56,4	150,4	104,9	133,1
Finlandia	80,9	49,8	188,7	657,0	55,2	50,8	134,6	17,8	25,9
Franca	153,3	100,7	230,6	380,2	54,2	54,3	126,7	42,7	120,1
...	...	...	...	...	...	...	...	...	...

Arquivo: Habitos_Alimentares (.txt)

FIA BUSINESS SCHOOL lab.data

Case: Hábitos Alimentares

Os dados abaixo são provenientes de uma pesquisa de consumo de alimentos em 25 países da Europa ao longo de determinado período. Ao todo, o consumo de nove grupos de alimentos foi analisado. Temos como objetivo agrupar os países que possuem comportamentos de alimentação semelhantes.

Adaptado a partir de: DASL *(The Data and Story Library*)

(a) Faça uma breve análise exploratória das variáveis de consumo alimentar.

(b) Padronize as variáveis.

(c) Calcule a matriz de distâncias euclidianas entre os 25 países.

(d) Faça uma análise de *cluster* utilizando o algoritmo hierárquico, com quatro diferentes critérios de ligação. A partir dos dendrogramas, qual método você sugere? E qual a quantidade de *clusters*?

(e) Segundo todos os critérios de ligação, qual o país que possui hábitos alimentares mais semelhantes aos da Itália? Qual o mais semelhante com a Romênia? E com a Holanda?

(f) Para o cenário escolhido no item (d), analise os hábitos alimentares dos países de cada *cluster.* Comente os resultados.

(g) Faça uma nova análise, agora utilizando o algoritmo *k*-médias, testando $k = 2, 3, 4$ e 5. A partir dos gráficos, qual a quantidade de *clusters* que você sugere?

(h) Para o cenário escolhido no item (g), analise os hábitos alimentares dos países de cada *cluster.* Comente os resultados.

(i) Quais as principais diferenças de interpretação entre as clusterizações escolhidas segundo o método hierárquico e segundo o método *k*-médias?

Arquivo: Habitos_Alimentares (.txt)

FIA BUSINESS SCHOOL lab.data

Case: *E-Commerce*

Uma varejista de *e-commerce* deseja realizar ações personalizadas de *cross sell* com os clientes que compraram apenas 1 produto eletrônico no site, em uma das seguintes categorias: telefone celular, televisão e computador. Para isso, deseja segmentar entre 4 e 6 perfis distintos de clientes, de acordo com a categoria de produto já comprado, valor gasto (R$) e nota de satisfação com a compra.

Variável	Descrição
Id_cliente	Código identificador do cliente
categoria	Categoria de eletrônico já adquirido pelo cliente: celular, televisão ou computador
valor_pago	Valor pago pelo cliente no produto eletrônico já adquirido, em R$
nota_satisfacao	Nota de satisfação com a compra de eletrônico realizada

Quantidade de clientes na base de dados: **2.992**

Arquivo: eCommerce (.txt)

Uma varejista de *e-commerce* deseja realizar ações personalizadas de *cross sell* com os clientes que compraram apenas 1 produto eletrônico no site, em uma das seguintes categorias: telefone celular, televisão e computador. Para isso, deseja segmentar entre 4 e 6 perfis distintos de clientes, de acordo com a categoria de produto já comprado, valor gasto (R$) e nota de satisfação com a compra.

(a) Faça uma breve análise exploratória das variáveis disponíveis.

(b) Calcule a matriz de distâncias de Gower entre todos os clientes.

(c) Faça uma análise de *cluster* utilizando o algoritmo *k*-medoides, com as três variáveis de interesse, para $k = 4$, 5 e 6. Analise descritivamente os *clusters* formados, para cada um dos três cenários. Comente os resultados.

(d) Qual dos três cenários você considera mais apropriado para a realização das ações de *cross sell*? Justifique.

Arquivo: eCommerce (.txt)

FIA BUSINESS SCHOOL lab.data

Referências Bibliográficas

ANÁLISE DE *CLUSTER*

- Härdle, W. K., Simar, L. *Applied Multivariate Statistical Analysis*. 4ª edição. Springer, 2014.
- Johnson, R. A., Wichern, D. W. *Applied Multivariate Statistical Analysis*. 6ª edição, Pearson Prentice-Hall Inc., 2007.
- Timm, N. H. *Applied Multivariate Analysis*. Springer-Verlag, 2002.
- Zelterman, D. *Applied Multivariate Statistics with R*. Springer, 2015.

lab.data

http://labdata.fia.com.br
Instagram: @labdatafia
Facebook: @LabdataFIA