
FIAP
MBA

# MBA em DATA SCIENCE & ARTIFICIAL INTELLIGENCE

## APPLIED STATISTICS

profregina.bernal@fiap.com.br
reginabernal@terra.com.br

## Dra. Regina Tomie Ivata Bernal
## Cientista de Dados na área da Saúde

Formação Acadêmica:

- Estatístico - UFSCar
- Mestre em Saúde Pública – FSP/USP
- Doutor em Ciências – Epidemiologia - FSP/USP

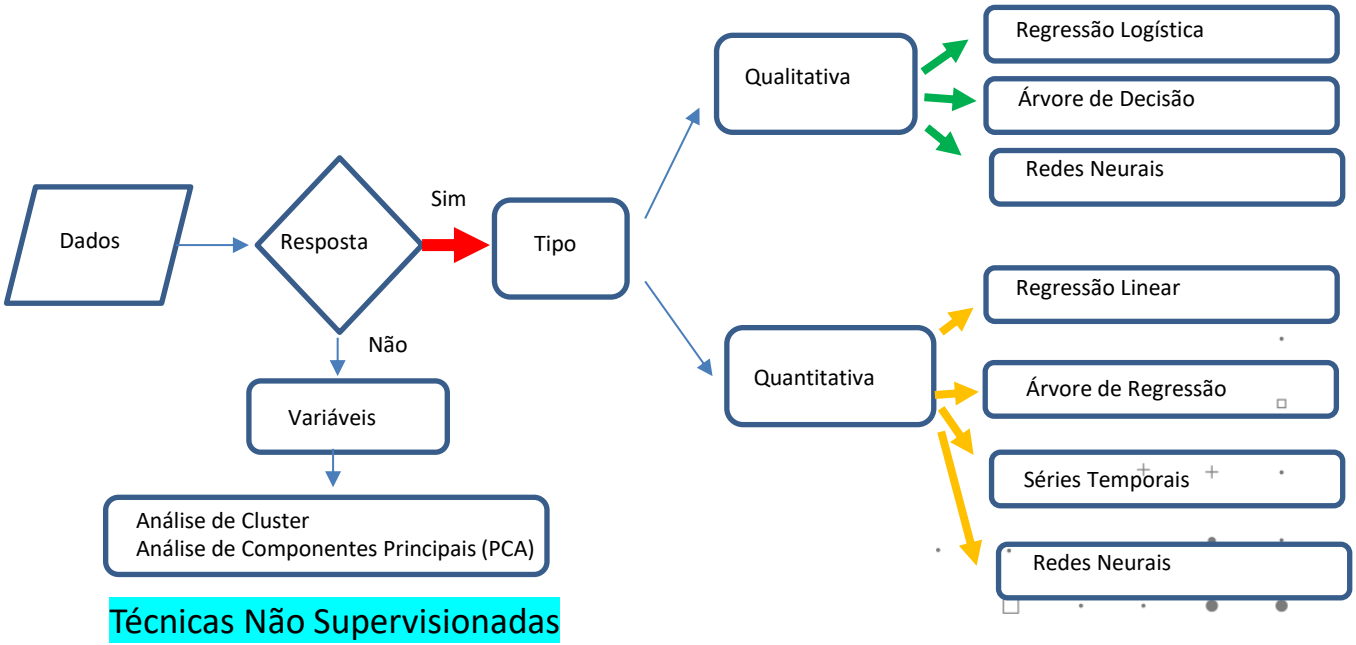
Atividades Profissionais:

- Professora de pós-graduação na FIAP
- Consultora externa da SVS/MS
- Cientista de Dados em Saúde

# Técnicas Estatísticas

**Extração de conhecimento em bases de dados**

**- A metodologia de data mining -**

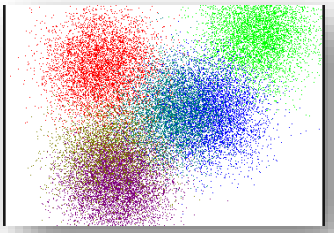


# TÉCNICAS NÃO SUPERVISIONADAS

# ANÁLISE ESTRUTURAL

## Análise de Conglomerados - Cluster

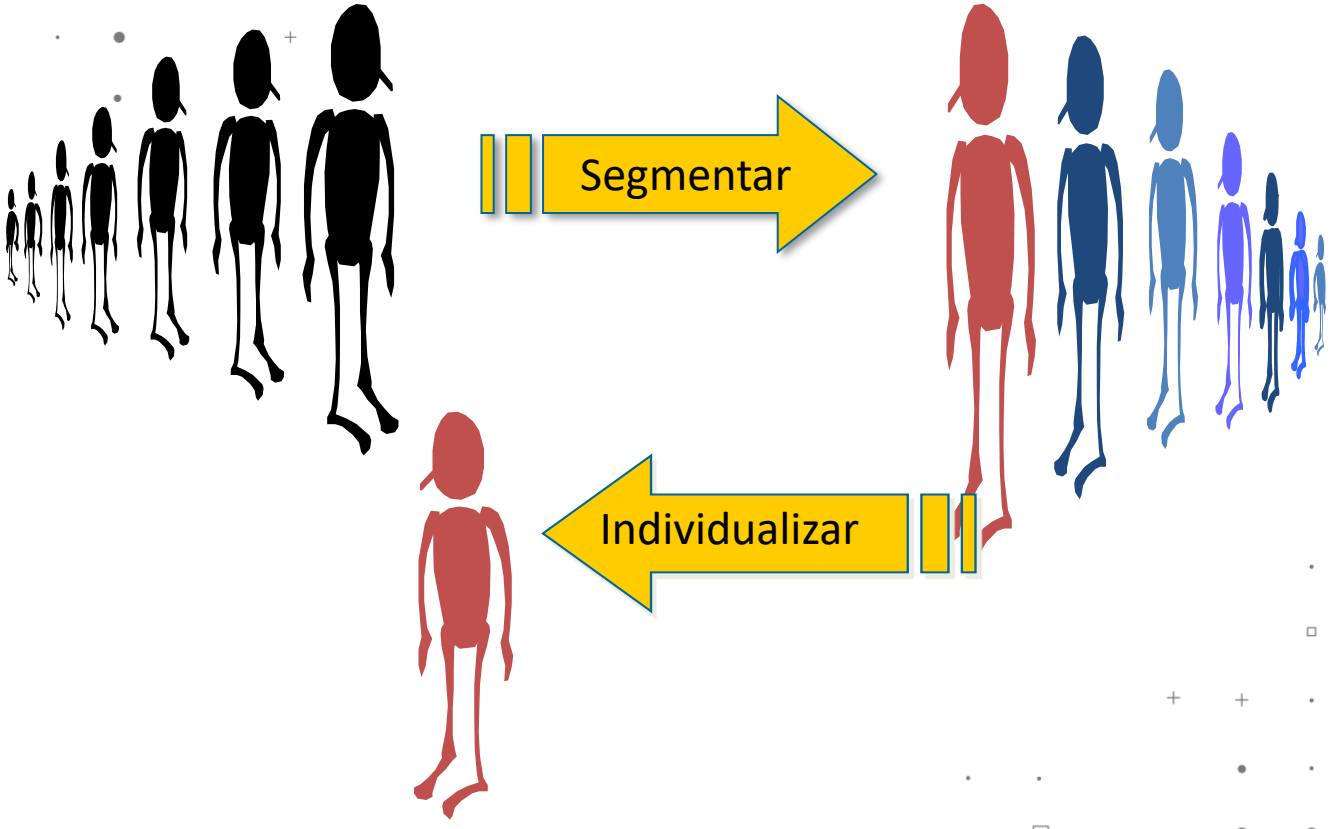
### Descobertas Não Supervisionadas de Relações



# Segmentação

A segmentação é um processo de agrupar clientes em grupos tais que apresentam características semelhantes entre os elementos do grupo e distintas entre os grupos.

# Instrumentação da Estratégia de Fidelização



# Tipos de Segmentações

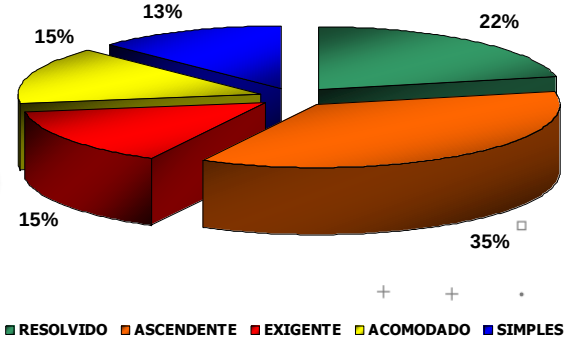
| | |
|---|---|
| Comportamental | Comportamento quanto ao uso do produto |
| Descritiva | Geo-Demográficos |
| Atitudinal | Valores, Hábitos e Atitudes do Cliente |
| Percepção | Considerações sobre o Produto |

**Conforme o objetivo selecionar a entidade de análise e as variáveis segmentadoras**

# Tipos de Segmentações

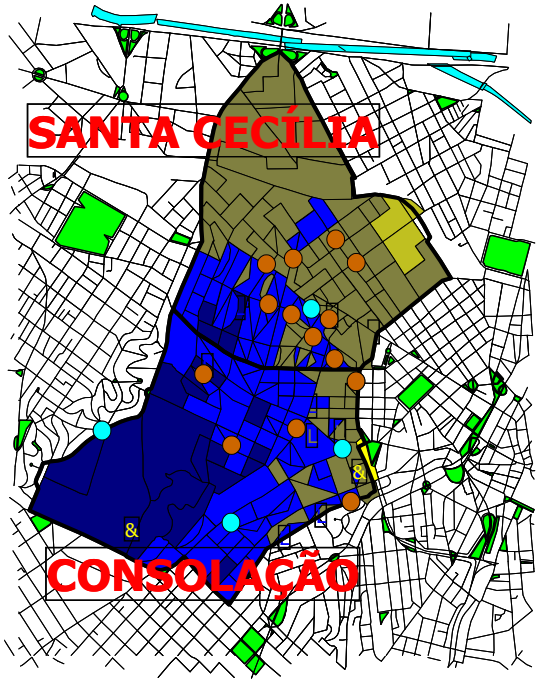
## Segmentação Comportamental de Clientes

- Cadastro
- Uso de voz / rádio
- Forma de pagamento
- aparelhos
- Serviços Contratados
- Uso de dados
- Comportamento de pagamento



# Tipos de Segmentações

## Modelos Geográficos




## Processo Interpretativo

**RENDA**

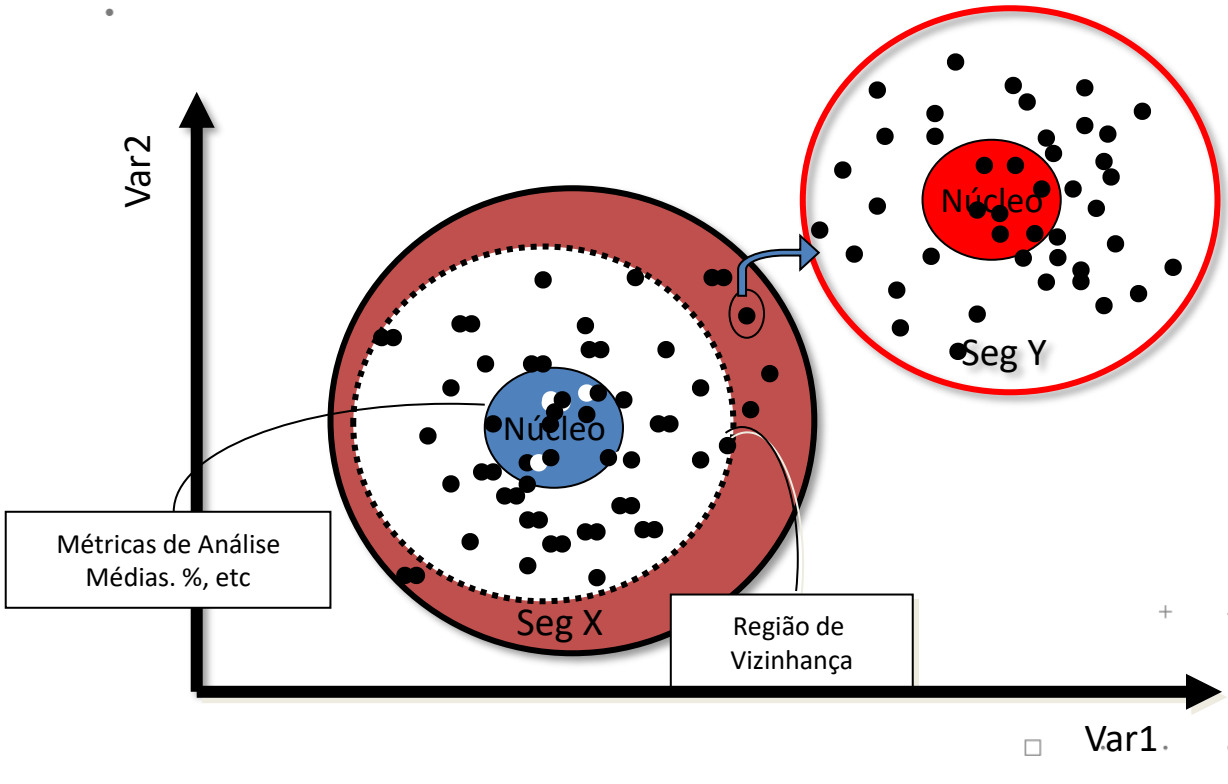
- Acima de 3
- 1,1 a 3,0
- 0,5 a 1,1
- 0, 3 a 0,5
- Abaixo de 0,3
- Inadimplentes
- Clientes

# Segmentação Geo-demográfica

## Município de São Paulo

- geo-marketing

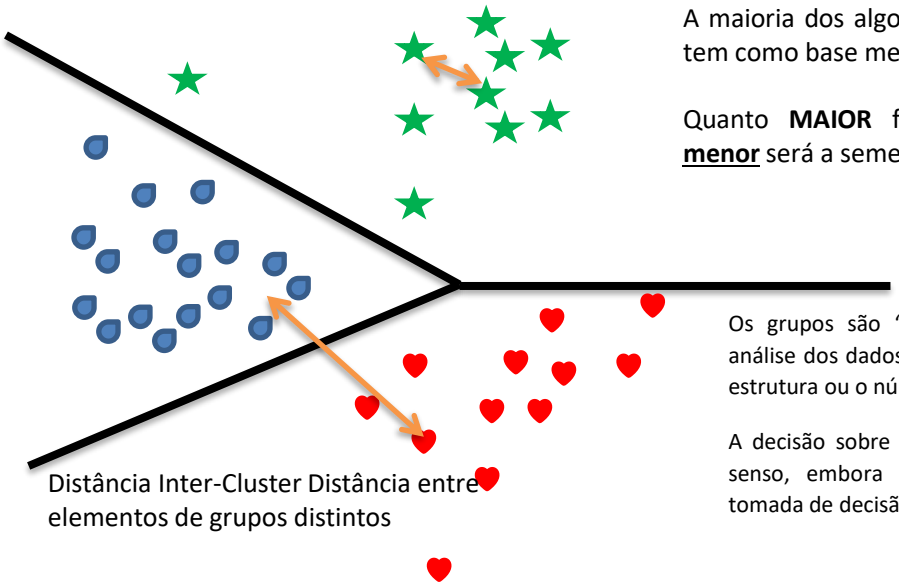

# Análise de Agrupamentos – Cluster Analysis



# Análise de Agrupamentos – Cluster Analysis

**Objetivo:** Separar um conjunto de objetos em grupos (clusters) de forma que os membros de qualquer grupo formado sejam os mais homogêneos possíveis com relação a algum critério ➔ **uso de medidas de distância**

Distância Intra-Cluster Distância entre elementos de um mesmo grupo.



Distância Inter-Cluster Distância entre elementos de grupos distintos

A maioria dos algoritmos de análise de agrupamento tem como base medidas de dissimilaridade:

Quanto **MAIOR** for a medida de dissimilaridade **menor** será a semelhança entre os indivíduos.

Os grupos são "naturais", isto é, surgem a partir da análise dos dados. Não existe suposição prévia sobre sua estrutura ou o número de grupos.

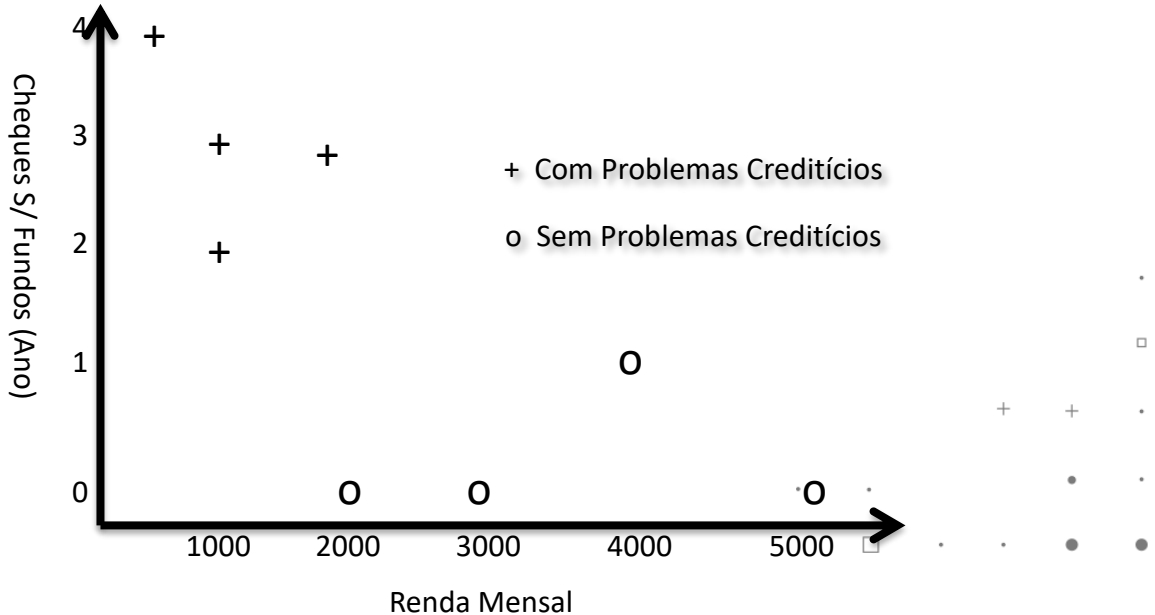
A decisão sobre o número de grupos depende de bom senso, embora existam critérios que dão suporte à tomada de decisão.

# Análise de Agrupamentos – Cluster Analysis

- **Elementos da Análise**

**Entidades** ← **Atributos**

- **Seleção Conjuntos de Atributos** - Variáveis Discriminantes



# Análise de Agrupamentos – Cluster Analysis

As etapas do processo de análise de clusters são:

1. Seleção da base de modelagem ➔ em função do objetivo (qual entidade, qual histórico..)
2. Seleção de atributos ➔ variáveis segmentadoras
3. Medida de proximidade
4. Critério de agrupamento
5. Algoritmo de agrupamento
6. Verificação dos resultados
7. Interpretação dos resultados

# Análise de Agrupamentos – Cluster Analysis

## Medidas de distância

Por exemplo a distância Euclidiana é calculada por:

$$d_{ij} = \sqrt{\left[\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2\right]}$$

Onde $x_{ik}$ é o valor da variável $X_k$ para o indivíduo (registro) *i* e $x_{jk}$ é o valor da mesma variável para o indivíduo *j.*

➔ Usualmente as variáveis são padronizadas antes de se calcular as distâncias, assim, as *p* variáveis serão igualmente importantes. Geralmente, a padronização feita é para que todas as variáveis (quantitativas) tenham média zero e variância 1.

# Análise de Agrupamentos – Cluster Analysis

**Padronização das variáveis :**

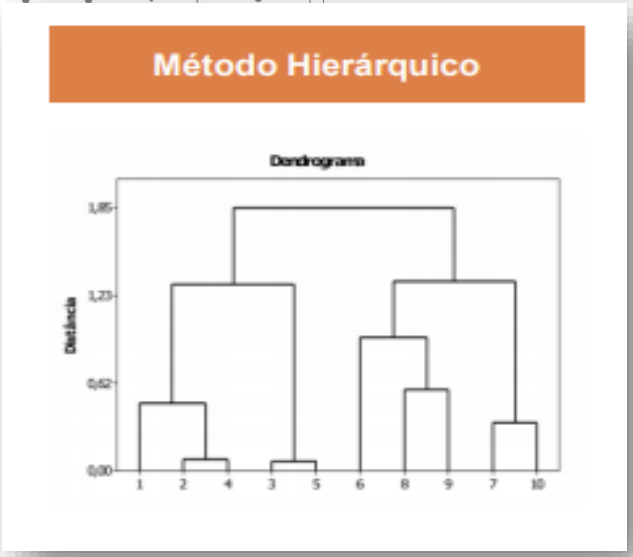
Os métodos baseados em distância são afetados pela diferença de escala entre os valores das variáveis/atributos, sendo necessário normalizar os atributos.

**Padronização** - Transforma os valores em números de desvios padrões a partir da média. É dada por: :

$$Z = \frac{X - \bar{X}}{S}$$

*Onde : $\bar{X}$ = Média da variável*
*S = desvio padrão*

# Metodologia: Classificação das Técnicas





**Hierárquicas** (envolvem a construção de uma hierarquia)

Aglomerativas

- ***todas as observações iniciam como sendo um grupo(unitário); grupos próximos são, então gradualmente juntados até, finalmente, todas as observações constituírem um único grupo.***

- Divisivas

- ***todas as observações iniciam num único grupo. Após são separados em dois grupos e assim por diante, até que cada observação seja o próprio grupo.***

**Não Hierárquicas** (trabalha com interações)

# Análise de Agrupamentos

## Métodos Hierárquicos de Agrupamentos:

## Método do Vizinho Mais Próximo

Método calcula-se a matriz de distâncias entre os “n” indivíduos da população, em seguida os indivíduos mais próximos são agrupados( método do encadeamento simples “single linkage method”

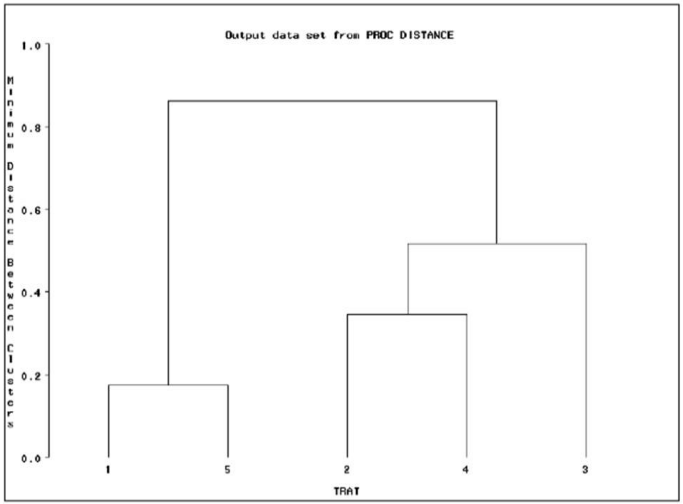
# Análise de Agrupamentos

## Métodos Hierárquicos de Agrupamentos:

## Exemplo de Agrupamento

- Método: vizinho mais próximo
- Dissimilaridade: distância euclidiana
- Dendrograma

Um *dendograma* é um meio prático de sumarizar um padrão de agrupamento. Este começa com todos os indivíduos separados (“folhas”) fundindo-se progressivamente em pares (folhas, ramos, galhos, tronco) até chegar a uma única raiz. A ordem dos indivíduos mostrada no dendograma e a ordem na qual os grupos entram no agrupamento.



# Análise de Agrupamentos

## Métodos Hierárquicos de Agrupamentos:

## ➔ Matriz de distância D1

Matriz de distância euclidiana entre os "n" indivíduos da população;

Como d (1 e 5) é a menor distância em D1, os indivíduos 1 e 5 são agrupados.

| Ind. (n) | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 5 | 10 | 7 | **1** |
| 2 | | 0 | 5 | 2 | 6 |
| 3 | | | 0 | 3 | 11 |
| 4 | | | | 0 | 8 |
| 5 | | | | | 0 |

Distância Euclidiana

$$d_{ij} = \sqrt{\left[\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2\right]}$$

# Análise de Agrupamentos

## Métodos Hierárquicos de Agrupamentos:

### ➔ Matriz de distância D2

Matriz de distância euclidiana entre d(1 e 5) e os demais indivíduos da população;

O menor valor em D2 é d(2 e 4)=2, então os indivíduos 2 e 4 são agrupados.

| | (15) | 2 | 3 | 4 |
|---|---|---|---|---|
| (15) | 0 | 5 | 10 | 7 |
| 2 | | 0 | 5 | **2** |
| 3 | | | 0 | 3 |
| 4 | | | | 0 |

# Análise de Agrupamentos

## Métodos Hierárquicos de Agrupamentos:

### ➔ Matriz de distância D3

Matriz de distância euclidiana entre d 2 e4 e os demais indivíduos da população;

O menor valor em D3 é d(2 e4)=3, então os indivíduo 3 é incluído no grupo 2 e 4.

| Ind. | (15) | (24) | 3 |
|---|---|---|---|
| (15) | 0 | 5 | 10 |
| (24) | | 0 | **3** |
| 3 | | | 0 |

### ➔ Matriz de distância D4

Matriz de distância euclidiana entre (234) e (15);

| | (15) | (234) |
|---|---|---|
| (15) | 0 | **5** |
| (234) | | 0 |

O grupo (234) é incluído no grupo(15), formando assim um único grupo .

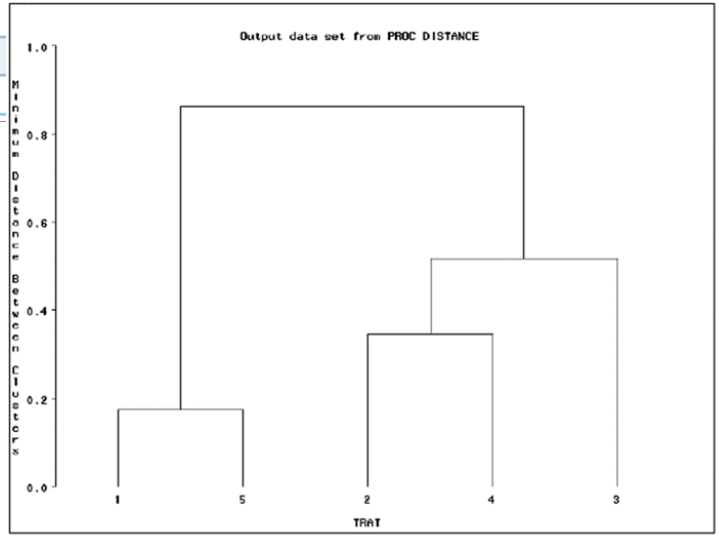
# Análise de Agrupamentos

## Métodos Hierárquicos de Agrupamentos:

## Resumo do método do vizinho mais próximo

➔ **Tabela resumindo passos, grupos e distâncias entre grupos.**

| PASSO | GRUPOS | DISTÂNCIA |
|---|---|---|
| 1 | 1,5 | 1 |
| 2 | 2,4 | 2 |
| 3 | 24,3 | 3 |
| 4 | 15,234 | 5 |



# Análise de Agrupamentos – Cluster Analysis

## Exemplo:

| Observação | Valor |
|---|---|
| a | 2 |
| b | 8 |
| c | 1 |
| d | 15 |
| e | 3 |
| f | 11 |
| g | 13 |
| h | 7 |
| i | 10 |

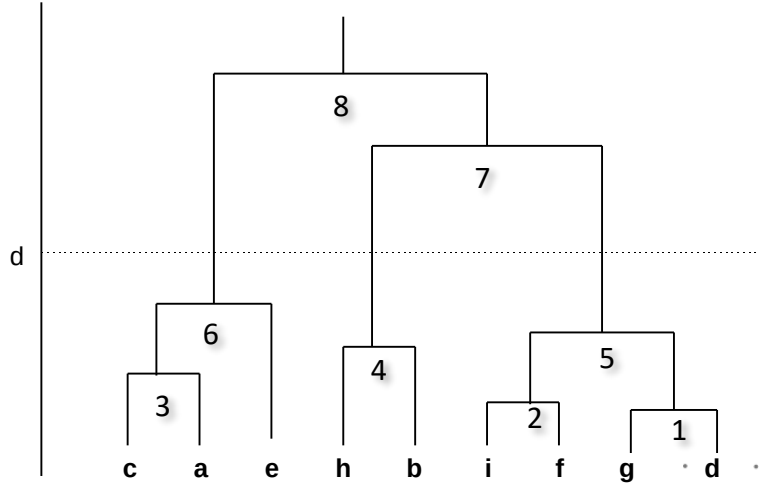
Distância=$|x_i-y_i|$

| | a | b | c | d | e | f | g | h | i |
|---|---|---|---|---|---|---|---|---|---|
| **a** | 0 | 6 | 1 | 13 | 1 | 9 | 11 | 5 | 8 |
| **b** | | 0 | 7 | 7 | 5 | 3 | 5 | 1 | 2 |
| **c** | | | 0 | 14 | 2 | 10 | 12 | 6 | 9 |
| **d** | | | | 0 | 12 | 4 | 2 | 8 | 5 |
| **e** | | | | | 0 | 8 | 10 | 4 | 7 |
| **f** | | | | | | 0 | 2 | 4 | 1 |
| **g** | | | | | | | 0 | 6 | 3 |
| **h** | | | | | | | | 0 | 3 |
| **i** | | | | | | | | | 0 |

# Análise de Agrupamentos – Cluster Analysis

## Técnicas Hierárquicas

Dendograma - Representação Gráfica de Agrupamento Aglomerativo



# Análise de Agrupamentos – Cluster Analysis

## Técnica Não-Hierárquica

**K-Means** - Uso intenso para grande volume de dados

- Parte de k sementes ou k clusters iniciais sobre os quais calcula as médias;
- Associa um item à semente/ média mais próxima (usando, por exemplo, a Distância Euclideana). Recalcula a média deste novo cluster e repete iterativamente esta etapa até que não haja mais realocação de elementos.

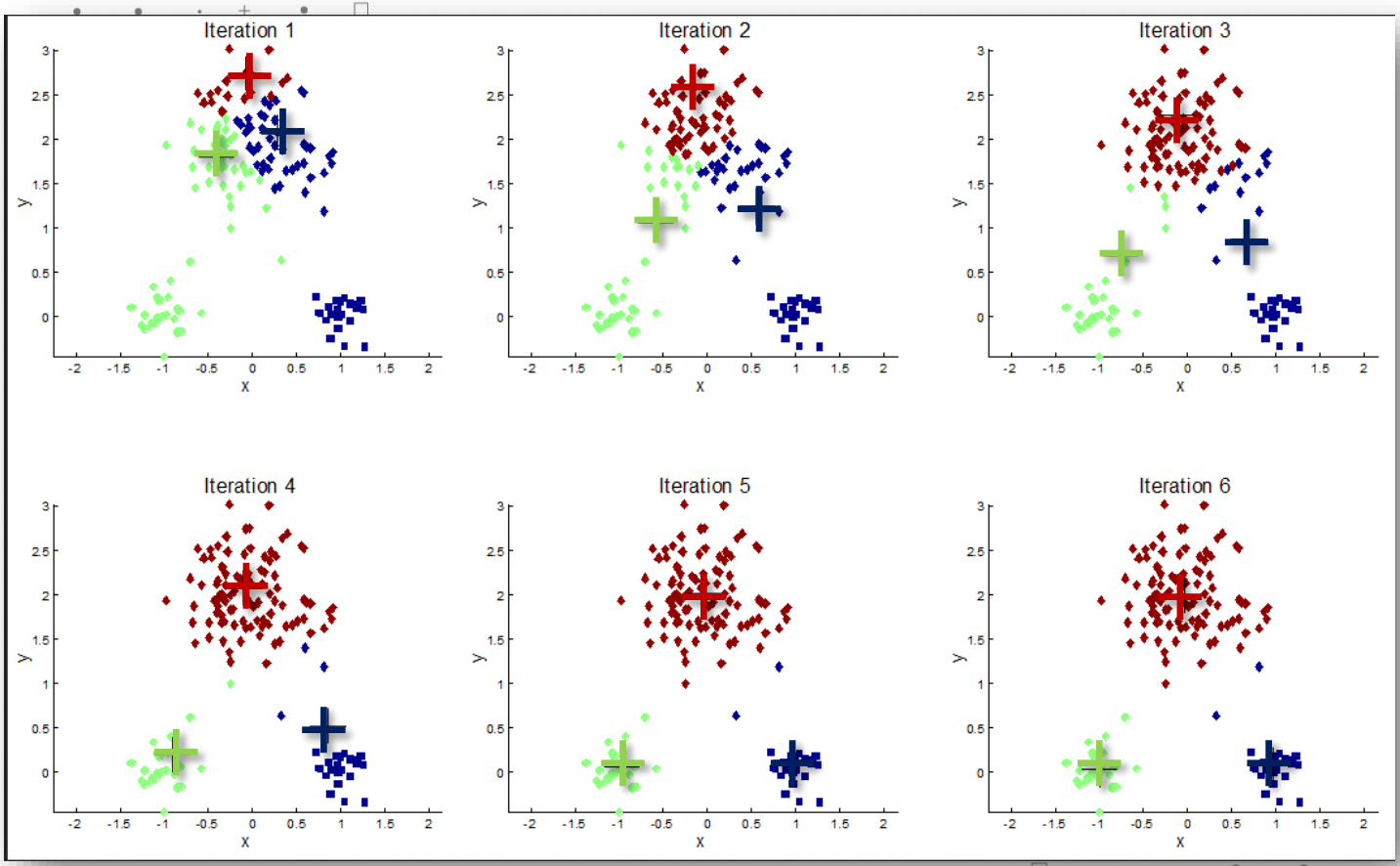
# Análise de Agrupamentos – Cluster Analysis

## Técnica Não-Hierárquica

**K-Means** - Uso intenso para grande volume de dados

- Parte de k sementes ou k clusters iniciais sobre os quais calcula as médias;
- Associa um item à semente/ média mais próxima (usando, por exemplo, a Distância Euclideana). Recalcula a média deste novo cluster e repete iterativamente esta etapa até que não haja mais realocação de elementos.

# ANÁLISE DE CONGLOMERADOS: Cluster Analysis - KMeans)



# Análise de Agrupamentos – Cluster Analysis

## Estatísticas a serem Avaliadas

- Número de Grupos
- Quantidade de Elementos no Grupo
- Média e Desvio-Padrão das Variáveis do Grupo
- Valor Máximo e Mínimo das Variáveis do Grupo
- Soma de Quadrados Médios dentro dos Grupos
- Soma de Quadrados Médios entre os Grupos

# Segmentação Comportamental
## Modelo RFV - Exemplo

Exemplo

Dados Internos

- Período da base de dados
  - Janeiro de 2.018 a Dezembro de 2.018 (1,7 MM clientes)

- Variáveis
  - Recência: Quantos dias atrás  última visita no site
  - Frequência: Quantos vezes por mês visita o site
  - Valor: Valor médio de compras em reais

- Técnica estatística: Análise de Cluster
  - Procedimento de aglomeração “K-Means”
  - Quantidade de Clusters: 4

# Análise de Cluster

## Exercitando!!!!!

## Segmentação RFV

# Segmentação Comportamental
# Modelo RFV - Resultados

Exemplo

## Perfil dos Segmentos

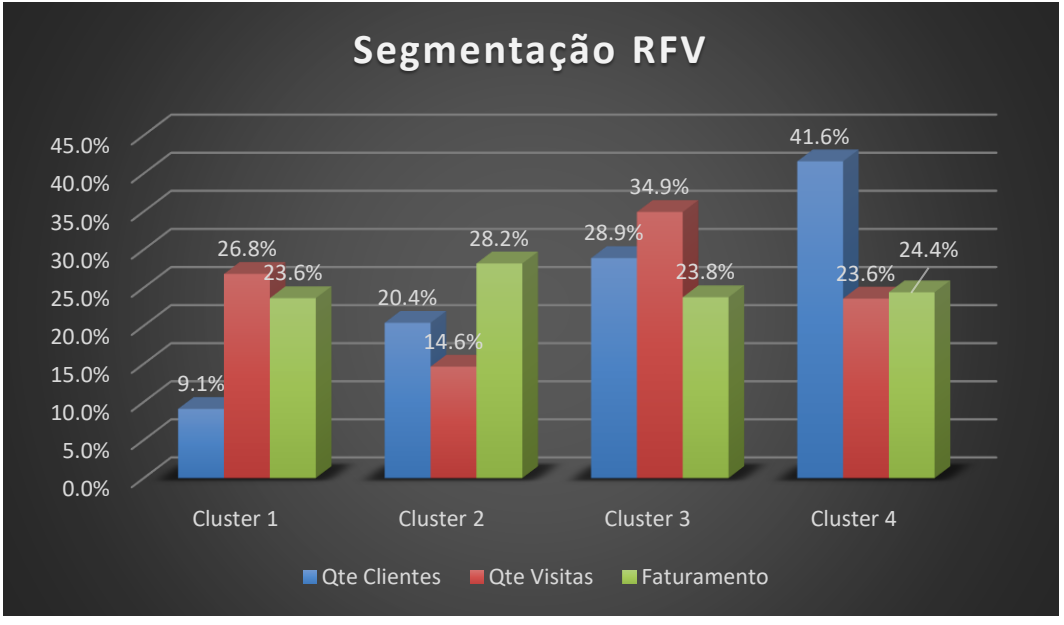
| Variáveis | Segmento 1 | Segmento 2 | Segmento 3 | Segmento 4 | Total |
|---|---|---|---|---|---|
| Média de visitas por mês | 7,8 | 1,9 | 3,2 | 1,5 | 2,6 |
| Recência em dias * | 3,4 | 9,3 | 6,7 | 15,0 | 10,4 |
| Valor médio por compra | R$ 490,47 | R$ 260,94 | R$ 155,21 | R$ 110,79 | R$ 188,81 |

*** Em média quantos dias atrás fez visita no site**

# Segmentação Comportamental
## Modelo RFV - Resultados

Distribuição da quantidade de Clientes, quantidade visitas e faturamento

Exemplo




**Fonte: Dados internos**

# DATA ANALYTICS

Capturar os dados

dados

Levantamento de dados

Cloud Services

Data Mart

Levantamento de dados

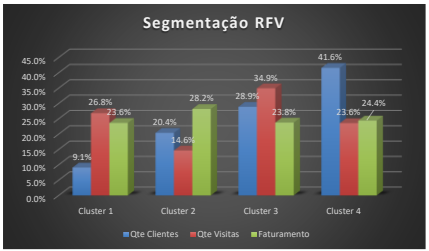
Data Mining

Estatística Descritiva

Amostragem
Regressão Linear
Séries Temporais

Ações

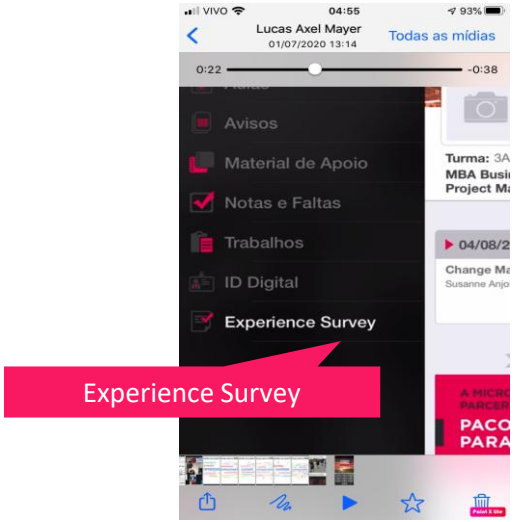
Cluster
Árvore Decisão
Regressão Logística
Regras de Associação

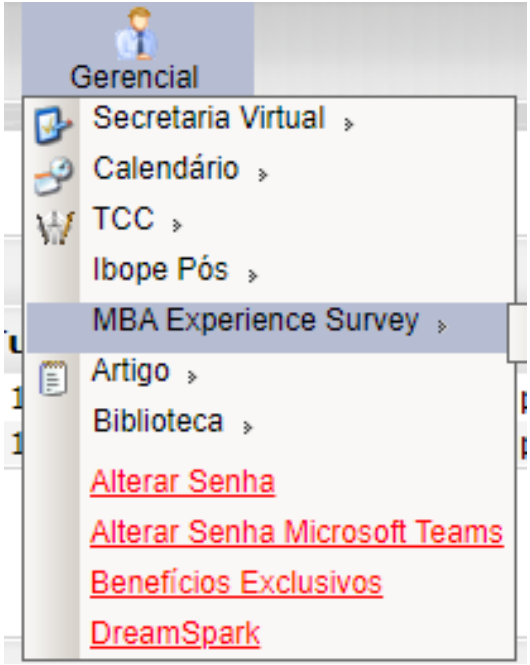


# O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)



# A grande finalidade do conhecimento não é conhecer, mas agir.

*T. Huxley*

# OBRIGADO

 / Regina T. I. Bernal


FIAP