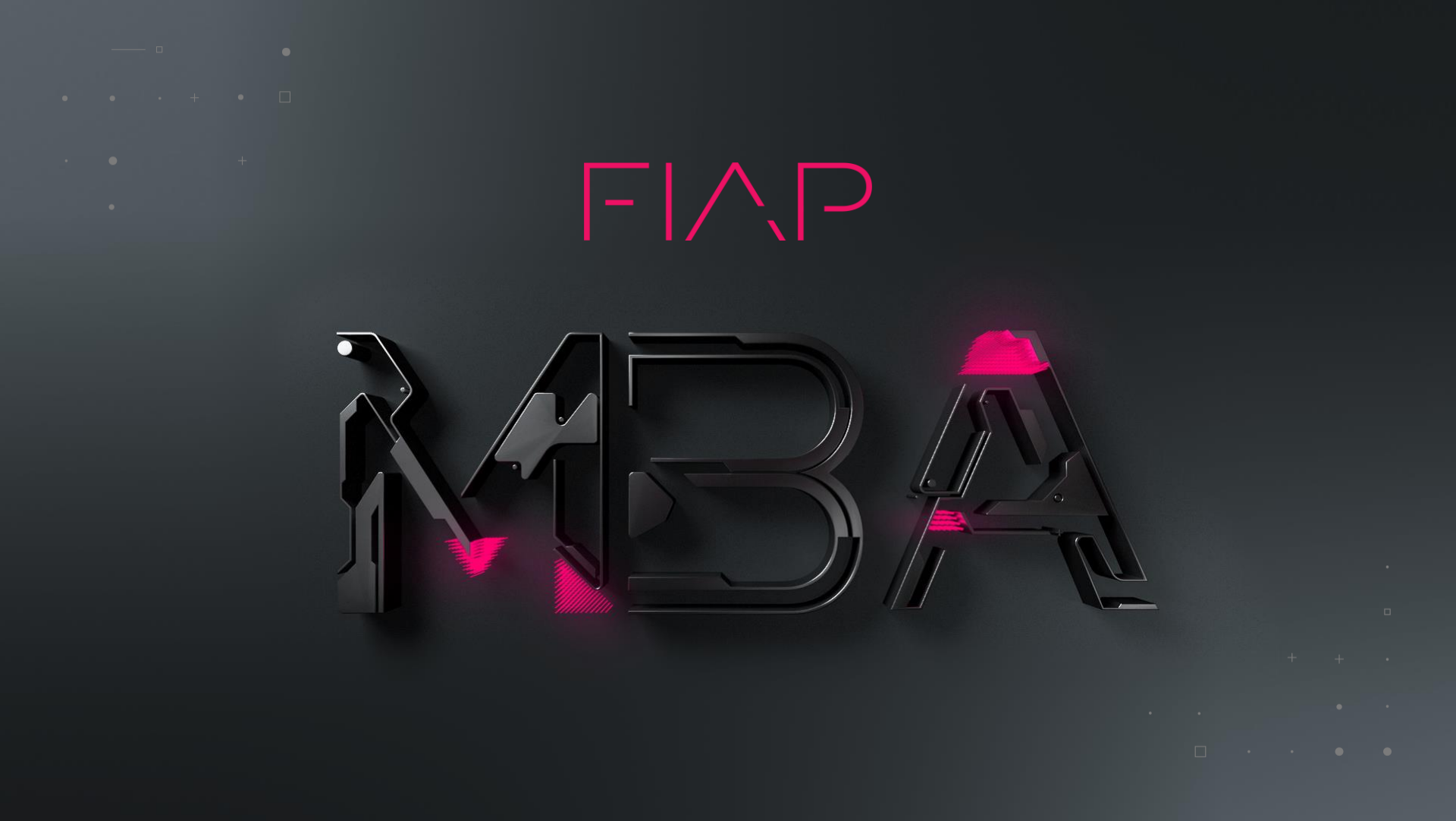

FIAP
MBA

# ADELAIDE ALVES DE OLIVEIRA

PROFESSORA

profadelaide.alves@fiap.com.br

## Formação Acadêmica

- Bacharel em Estatística – UNICAMP
- Mestre em Ciências – FSP/USP

## Atividades Profissionais

- Diretora Técnica Estatística da empresa **SD&W - www.sdw.com.br**
- Professora de Fundamentos Estatísticos, DataMining, Análise Preditiva e Machine Learning na FIAP dos cursos MBA: Big Data, Data Science, Business Intelligence & Analytics, Digital Data Marketing, IA & ML e Engenharia de Dados e nos Shift: People Analytics e Python Journey

# Conceitos Estatísticos para IA

# Introdução...(Voltando

➔ ... enfim, seus dados não servem para nada até que você saiba como tirar informações deles

| DESCRITIVO<br>O que aconteceu? | DIAGNÓSTICO<br>Por que isto aconteceu? | PREDITIVO<br>O que acontecerá? | PRESCRITIVO<br>O que posso fazer? |
|---|---|---|---|
| Quantos clientes temos cancelados voluntariamente?<br>Quais os tipos de produtos?<br>Qual a região que mora?<br>Qual o tempo é cliente da empresa? | Qual a relação entre o cancelamento e região?<br>Quais os motivos de cancelamentos.<br>Qual a taxa de cancelamento por safra? | Qual a probabilidade de um cliente cancelar ?<br>Quais são os clientes que que queremos reter?<br>Qual é a segmentação de valor dos clientes de maior propensão ao cancelamento? | Lista de ações para reter o cliente que está ativo e que tem a propensão ao cancelamento a um determinado tempo. Qual o canal que vamos utilizar para cada cliente? Que ações? |

# ANÁLISE MULTIVARIADA

## Análise Exploratória dos Dados

### Análise de Discriminação de Estrutura

- Técnicas de dependência.
- Técnicas Multivariadas aplicáveis quando uma das variáveis **pode ser identificada como dependente (variável target)**, e as restantes como variáveis independentes.

**Análise Supervisionada**

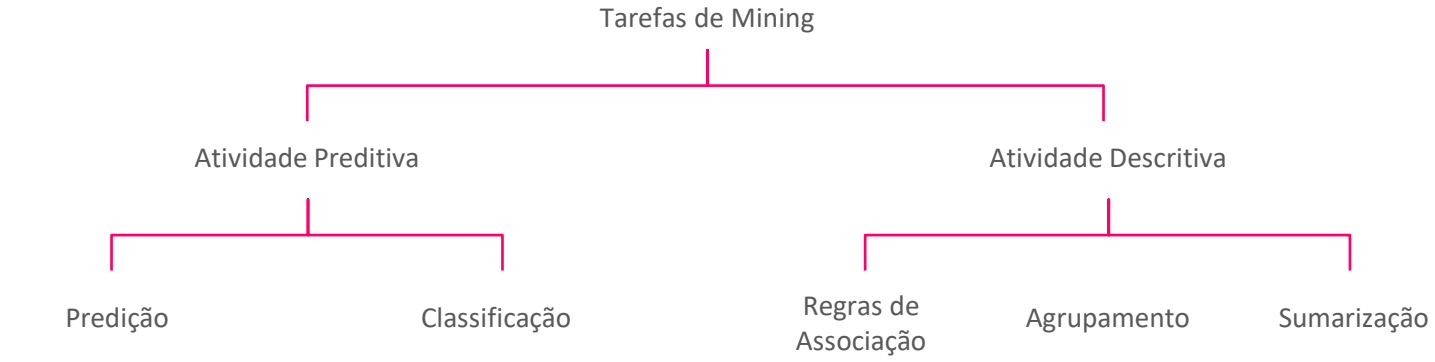
### Análise Estrutural

- Técnicas de Interdependência.
- Técnicas Multivariadas que procuram agrupar dados com base em semelhança, permitindo assim a interpretação das estruturas dos dados. **Não há distinção entre variáveis dependentes e independentes**.

**Análise Não Supervisionada**

# DATA MINING

Aplicações práticas de Data Mining podem ser categorizadas de acordo com a tarefa que se pretende resolver.



# ANÁLISE ESTRUTURAL

## ANÁLISE DE CONGLOMERADOS – CLUSTER

Descobertas **Não** Supervisionadas de Relações

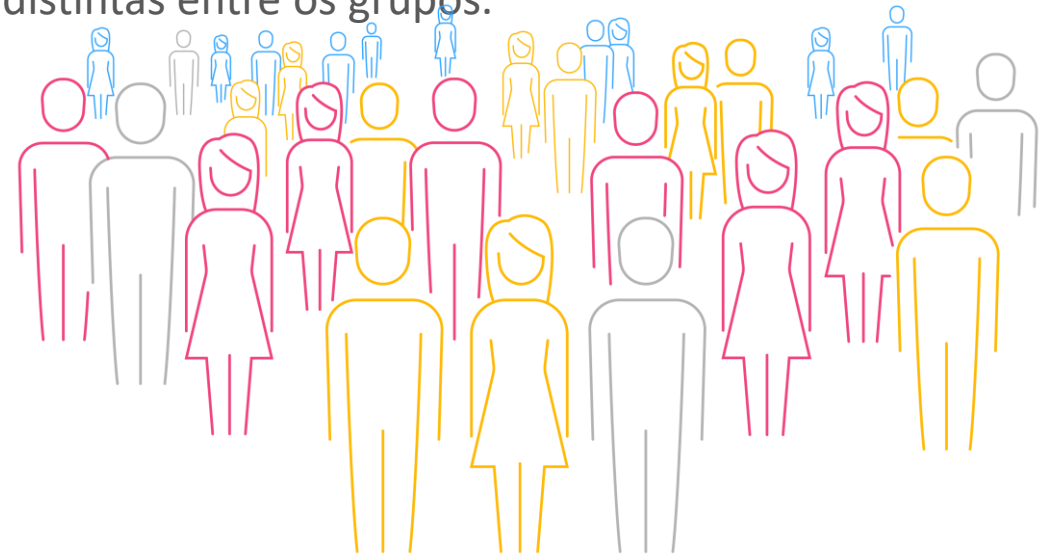
# ANÁLISE ESTRUTURAL

## ANÁLISE DE CONGLOMERADOS

## CLUSTER ANALYSIS

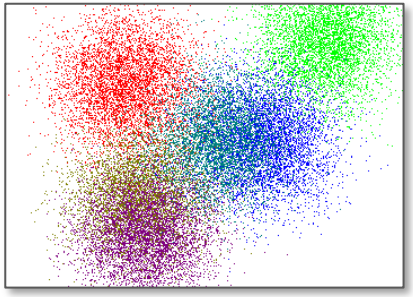
# SEGMENTAÇÃO

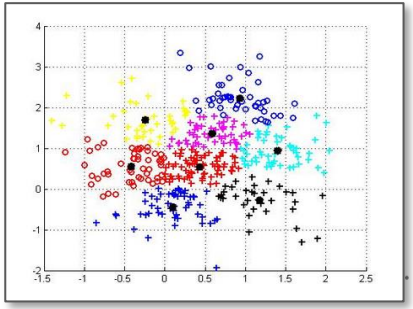
A segmentação é um processo de agrupar clientes em grupos tais que apresentam características semelhantes entre os elementos do grupo e distintas entre os grupos.



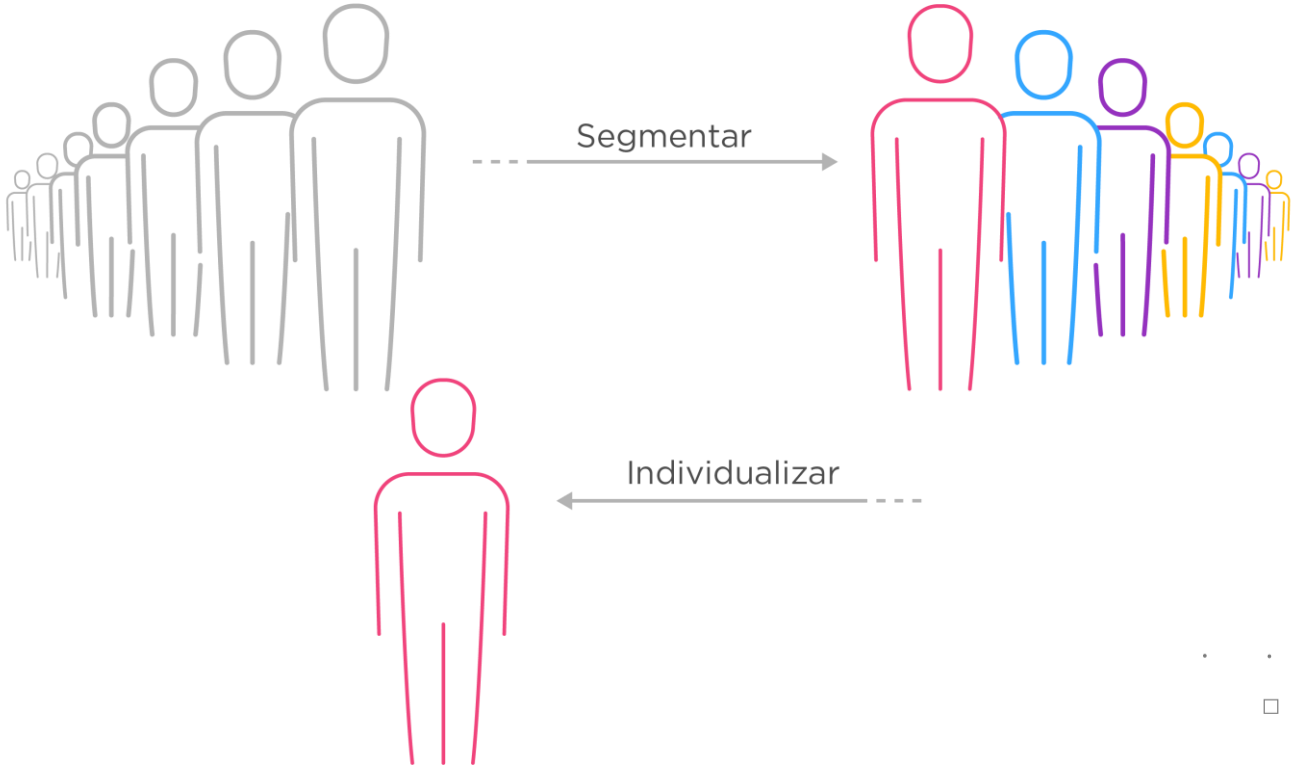
# TÉCNICA DE AGRUPAMENTO:
# ANÁLISE DE CLUSTERS

- Agrupamento (Clusterização): Consiste em segmentar os registros do conjunto de dados em subconjuntos ou clusters, de tal forma que os elementos de um cluster compartilhem propriedades comuns que os distingam de elementos nos demais clusters. O objetivo nesta tarefa é maximizar a similaridade intracluster e minimizar a similaridade intercluster.





# INSTRUMENTAÇÃO DA ESTRATÉGIA DE FIDELIZAÇÃO



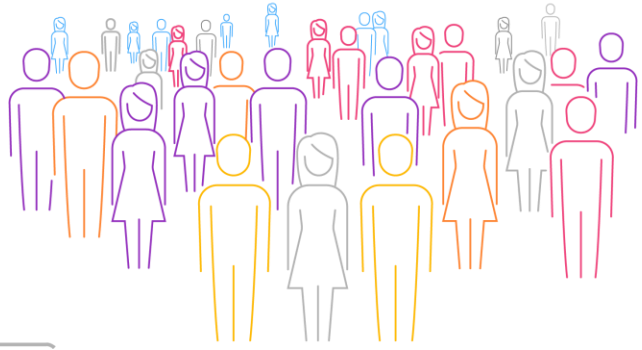
# TIPOS DE SEGMENTAÇÕES

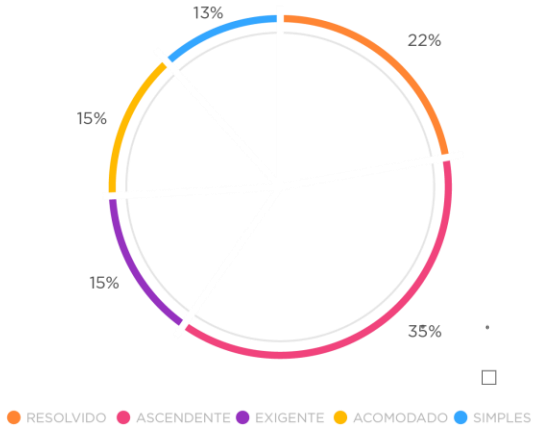
| | |
|---|---|
| Comportamental | Comportamento quanto ao uso do produto |
| Descritiva | Geodemográficos |
| Atitudinal | Valores, Hábitos e Atitudes do Cliente |
| Percepção | Considerações sobre o Produto |

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

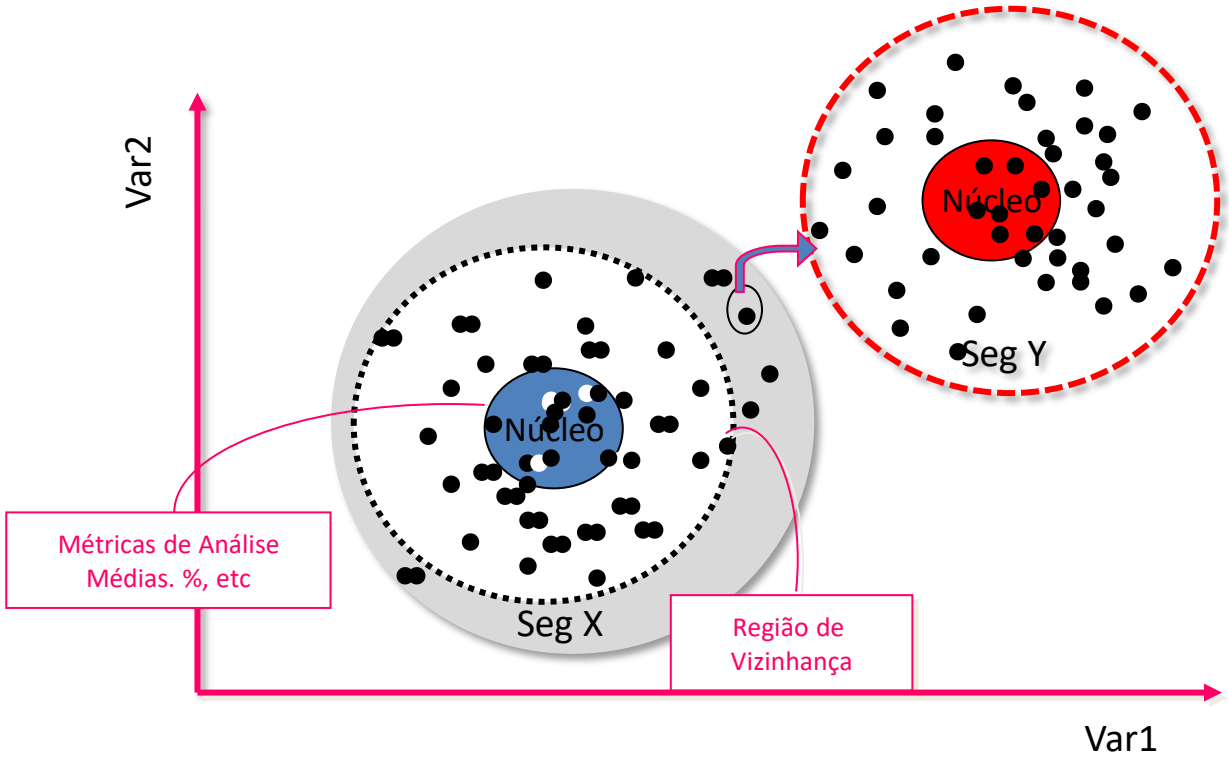
# TIPOS DE SEGMENTAÇÕES

## Segmentação Comportamental do Cliente

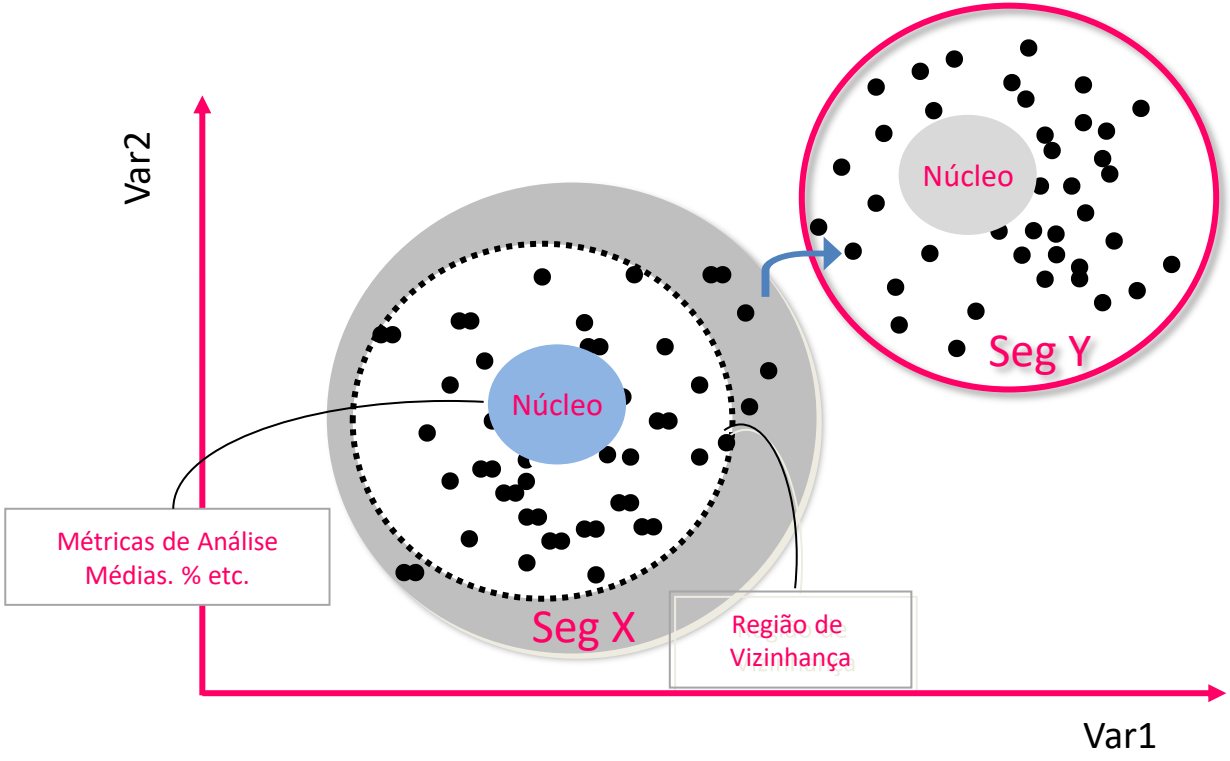


# Análise de Agrupamentos – Cluster Analysis



# ANÁLISE DE AGRUPAMENTOS
## CLUSTER ANALYSIS



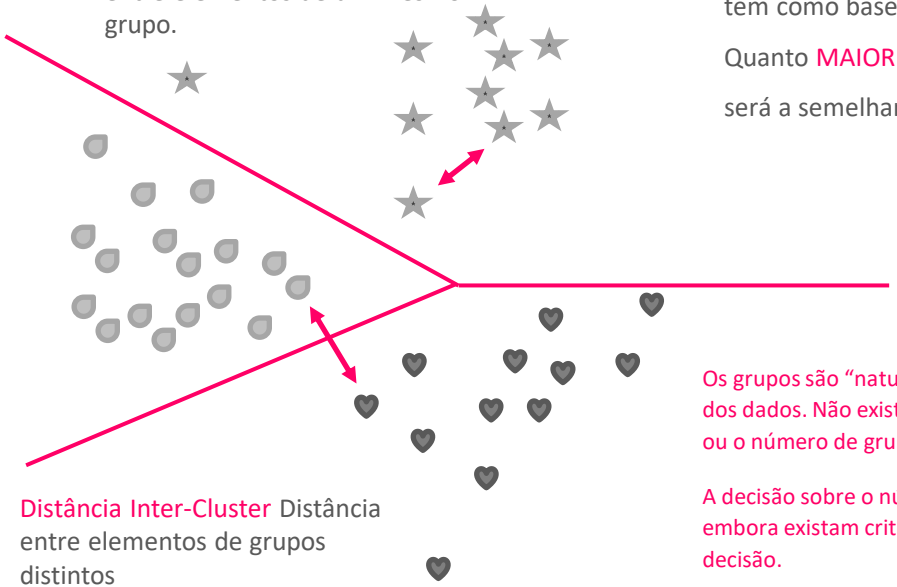
# ANÁLISE DE AGRUPAMENTOS
# CLUSTER ANALYSIS

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

# ANÁLISE DE AGRUPAMENTOS
## CLUSTER ANALYSIS

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.

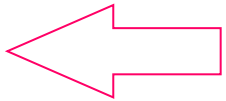
Os grupos são “naturais”, isto é, surgem a partir da análise dos dados. Não existe suposição prévia sobre sua estrutura ou o número de grupos.

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

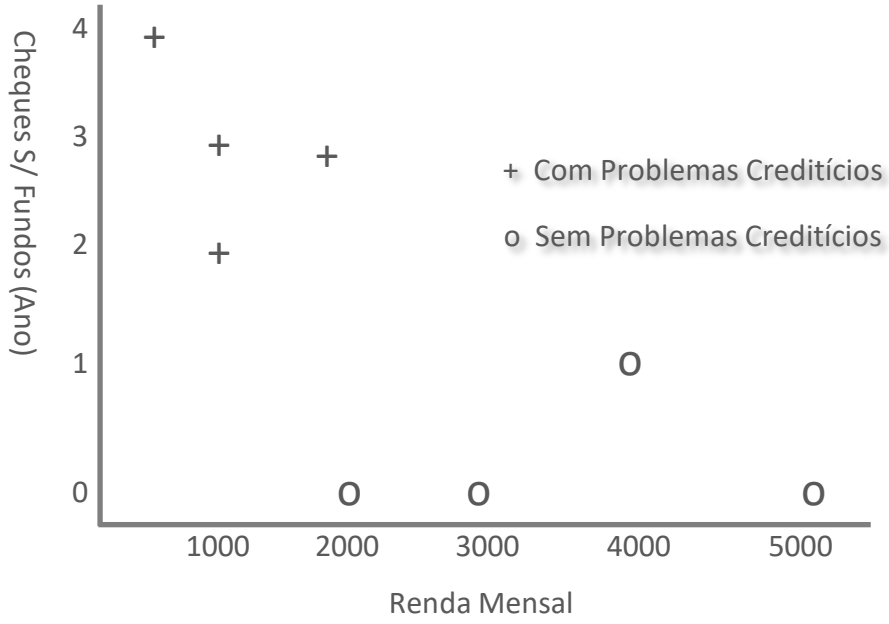
Entidades



Atributos

# ANÁLISE DE AGRUPAMENTOS
## CLUSTER ANALYSIS

- Seleção Conjuntos de Atributos - Variáveis Discriminantes e não colineares



# 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*

# 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.

Base Original

| id | Salário | Idade |
|---|---|---|
| 1 | 16.284 | 47 |
| 2 | 3.500 | 22 |
| 3 | 13.751 | 24 |
| 4 | 4.751 | 24 |
| 5 | 6.751 | 25 |
| 6 | 8.750 | 26 |

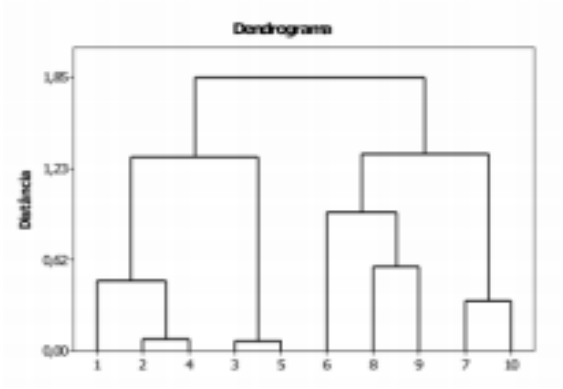
| | Salário | Idade |
|---|---|---|
| Média | 8.539,5 | 31,1 |
| Desvio | 4.716,4 | 9,3 |

Base com Padronização da Variáveis

| id | Salário | idade |
|---|---|---|
| 1 | 1,64 | 1,71 |
| 2 | -1,07 | -0,97 |
| 3 | 1,10 | -0,76 |
| 4 | -0,80 | -0,76 |
| 5 | -0,38 | -0,65 |
| 6 | 0,04 | -0,54 |

# ANÁLISE DE AGRUPAMENTOS
## CLASSIFICAÇÃO DAS TÉCNICAS

**Método Hierárquico**




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.

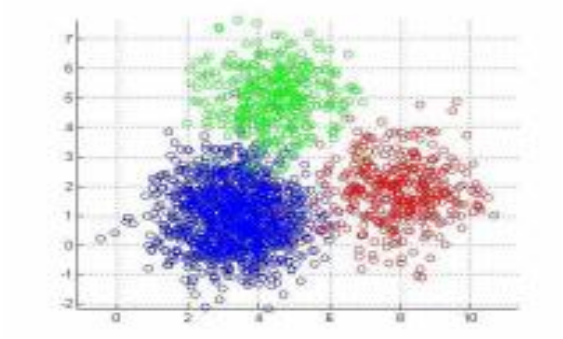
Divisivas

- 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.

# ANÁLISE DE AGRUPAMENTOS

## CLASSIFICAÇÃO DAS TÉCNICAS

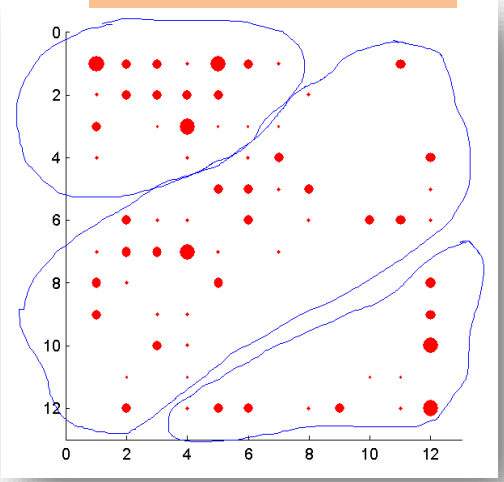
**Método Não-Hierárquico**

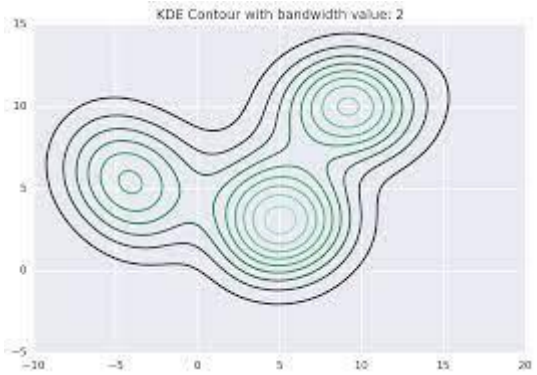


Não Hierárquicas (trabalha com interações)

# ANÁLISE DE AGRUPAMENTOS
## CLASSIFICAÇÃO DAS TÉCNICAS

Método por densidade



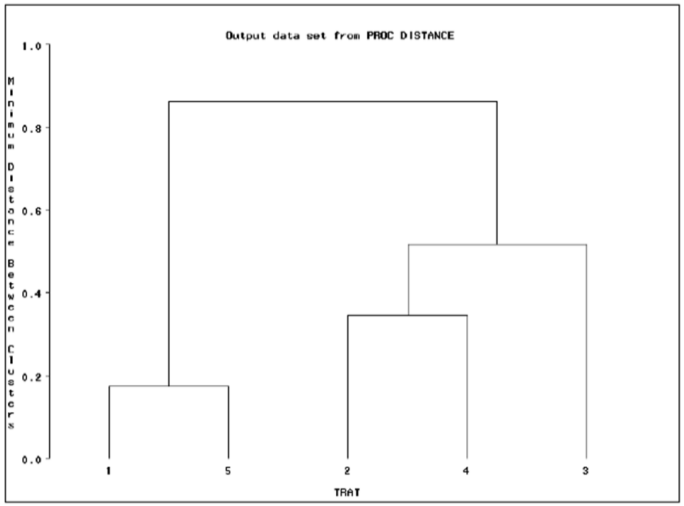

Busca existência de regiões densas de dados, separadas por regiões com baixa densidade de dados.

# ANÁLISE DE AGRUPAMENTOS
## MÉTODO HIERÁRQUICO

Métodos Hierárquicos de Agrupamentos:

Exemplo de Agrupamento

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



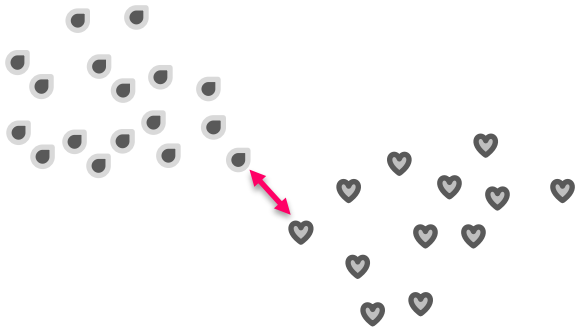

Um dendograma é um meio prático de sumarizar um padrão de agrupamento. Ele começa com todos os indivíduos separados (“folhas”) fundindo-se progressivamente em pares (folhas, ramos, galhos, tronco) até chegar a uma única raiz. A ordem dos indivíduos mostrada no dendograma é a ordem na qual os grupos entram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

# ANÁLISE DE AGRUPAMENTOS
## MÉTODO HIERÁRQUICO

Métodos Hierárquicos de Agrupamentos:

- Método do vizinho mais próximo

Método calcula 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

## MÉTODO HIERÁRQUICO

Métodos Hierárquicos de Agrupamentos:

- Matriz de distância D1

Matriz de distância euclidiana entre os “n” indivíduos da população;

Como d(15) 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 |

Exemplo

# ANÁLISE DE AGRUPAMENTOS

## MÉTODO HIERÁRQUICO

Métodos Hierárquicos de Agrupamentos:

- Matriz de distância D2

Matriz de distância euclidiana entre d(15) e os demais indivíduos da população;

O menor valor em D2 é d(24)=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 |

Exemplo

# ANÁLISE DE AGRUPAMENTOS

## MÉTODO HIERÁRQUICO

Métodos Hierárquicos de Agrupamentos:

- Matriz de distância D3

Matriz de distância euclidiana entre d(24) e os demais indivíduos da população;

O menor valor em D3 é d(24)=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 |

Exemplo

# ANÁLISE DE AGRUPAMENTOS

## MÉTODO HIERÁRQUICO

Métodos Hierárquicos de Agrupamentos:

- Matriz de distância D4

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

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

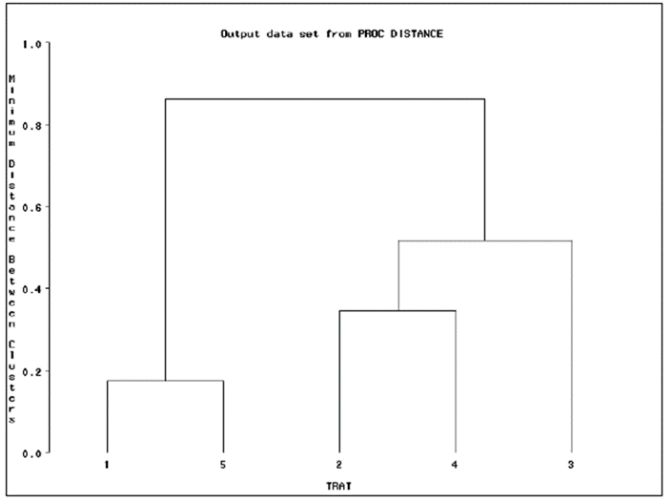
O grupo (234) é incluído no grupo (15), formando assim um único grupo .

Exemplo

# ANÁLISE DE AGRUPAMENTOS
## MÉTODO HIERÁRQUICO

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 |

Exemplo

# ANÁLISE DE AGRUPAMENTOS
## MÉTODO HIERÁRQUICO

| 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 |

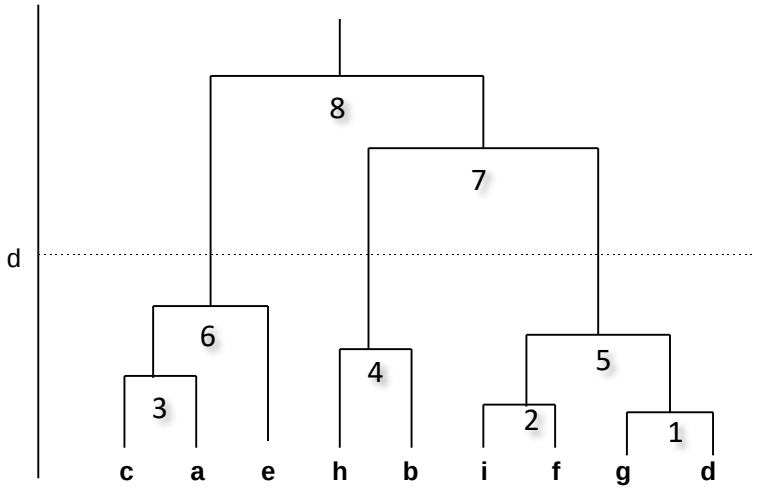
Exemplo

# ANÁLISE DE AGRUPAMENTOS

## MÉTODO HIERÁRQUICO

Técnicas Hierárquicas

- Dendograma – Representação Gráfica de Agrupamento Aglomerativo




Exemplo

# ANÁLISE DE AGRUPAMENTOS
## MÉTODO NÃO HIERÁRQUICO

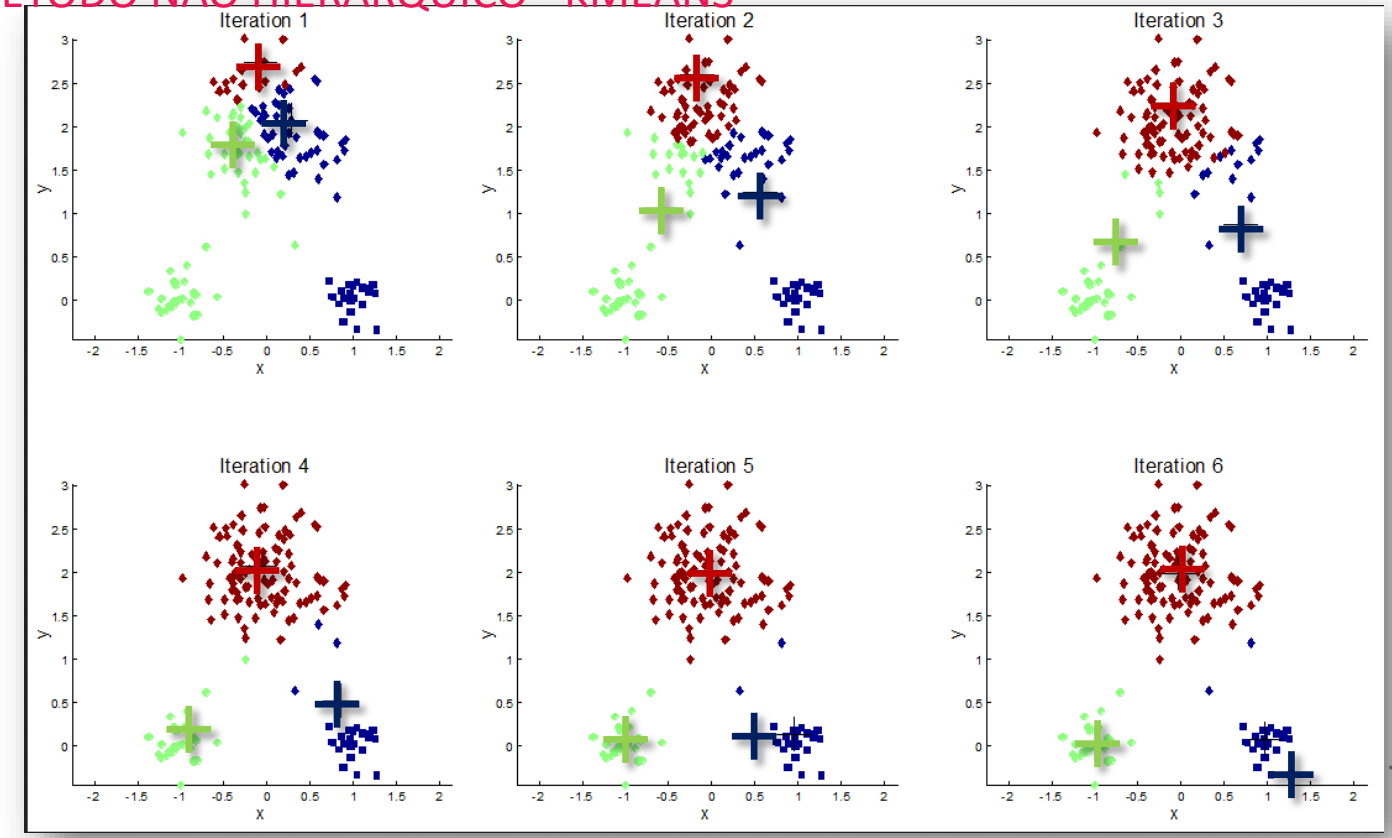
Técnica Não-Hierárquica

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 desse novo cluster e repete iterativamente essa etapa até que não haja mais realocação de elementos.

# ANÁLISE DE AGRUPAMENTOS

## MÉTODO NÃO HIERÁRQUICO - 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

Dados Internos

- Período da base de dados
  - Janeiro de 2.019 a Dezembro de 2.019 (1,85 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

Exemplo

# SEGMENTAÇÃO COMPORTAMENTAL

## MODELO RFV – RESULTADOS

### 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 |

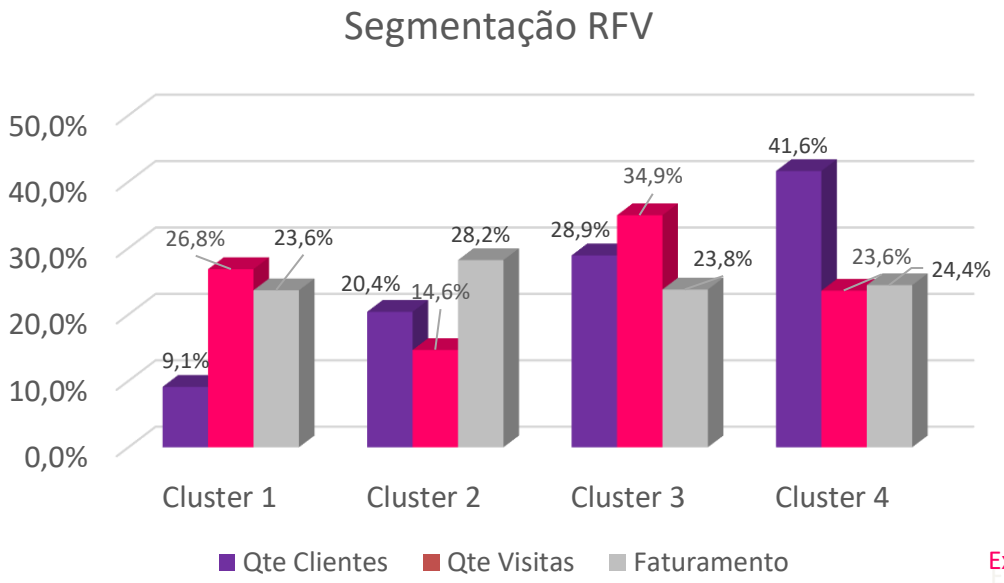
Exemplo

*** 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

# ANÁLISE DE AGRUPAMENTOS
## MÉTODO POR DENSIDADE

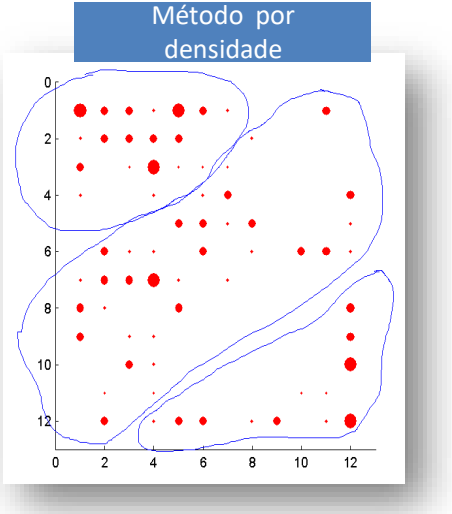
Os métodos de agrupamento baseados em densidade tentam suprir a necessidade de métodos capazes de descobrir grupos com formas arbitrárias. Nestes algoritmos, a ideia de grupos é baseada **na existência de regiões densas de dados, separadas por regiões com baixa densidade de dados.**

Alguns exemplos de algoritmos desta classe são:

DBSCAN: Density-Based Spatial Clustering of Applications with Noise;

OPTICS: Ordering Points to Identify the Clustering Structure;

DENCLUE: Density-based Clustering

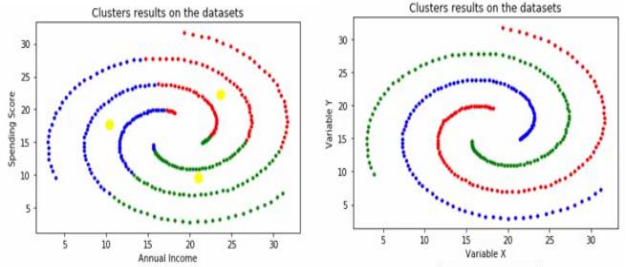


# ANÁLISE DE AGRUPAMENTOS
## MÉTODO POR DENSIDADE

**DBSCAN**: o processo executado no algoritmo “encontra” regiões com densidade suficientemente alta para descobrir os clusters, considerando um conjunto de dados “com ruído”.

Neste algoritmo, utilizado em *Machine Learning* para clusterização das observações utilizando medida de distância no espaço. É uma técnica não paramétrica baseada na densidade. Um cluster é definido como o um conjunto máximo de density-connected points. Um cluster baseado em densidade é um conjunto de objetos conectados “por densidade” que é máximo com respeito a densidade alcançável. Todo objeto não contido em um cluster é considerado ruído.

**Exemplo:**

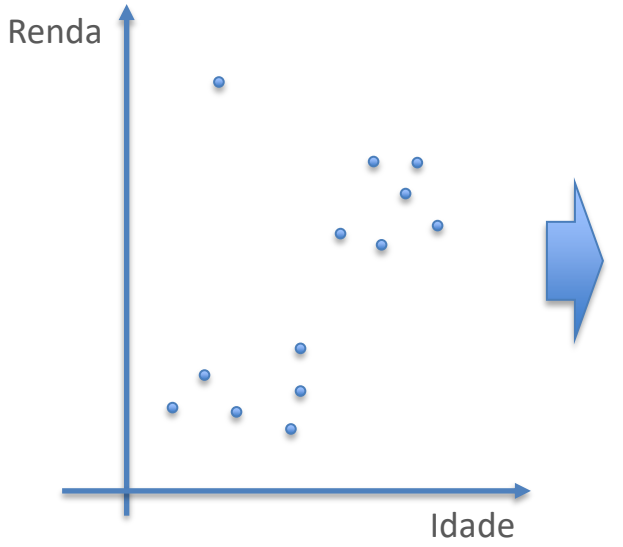


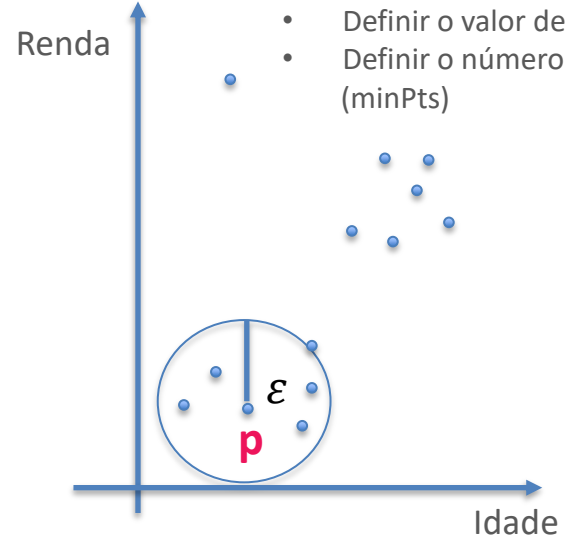

KMeans DBSCAN

*Fonte:*
*Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, №34, pp. 226–231)*

# ANÁLISE DE AGRUPAMENTOS
## MÉTODO POR DENSIDADE

### DBSCAN - Como funciona?

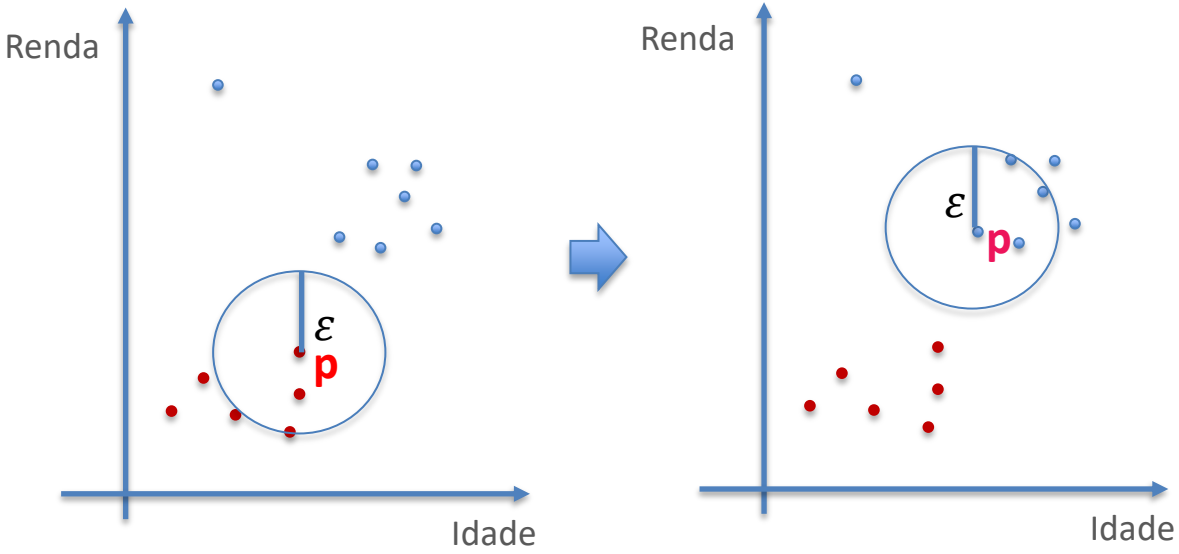



- Seleciona um ponto aleatoriamente
- A vizinhança do objeto p com raio ε é chamada de ε-vizinhança de p
- Definir o valor de $\varepsilon$
- Definir o número mínimo de pontos (minPts)

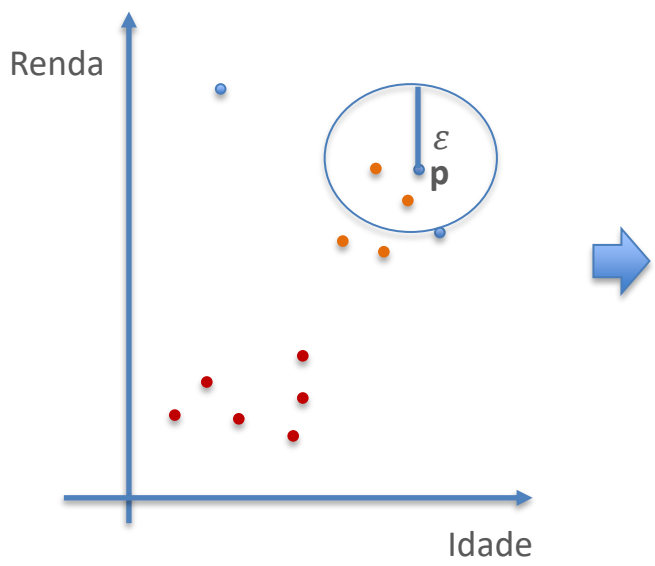
# ANÁLISE DE AGRUPAMENTOS
## MÉTODO POR DENSIDADE

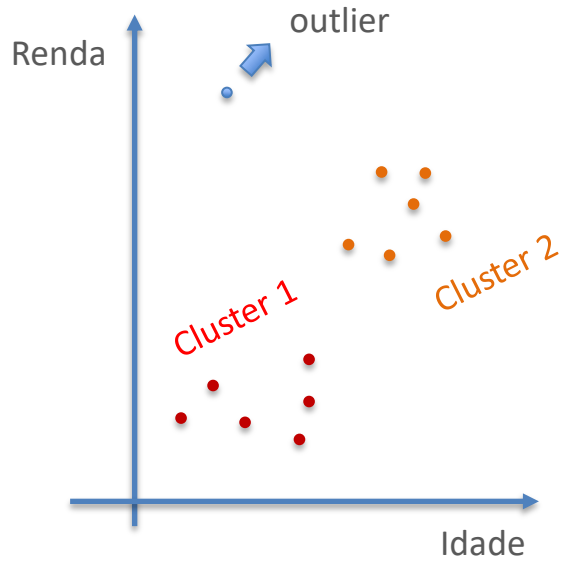
DBSCAN - Como funciona?



# ANÁLISE DE AGRUPAMENTOS
## MÉTODO POR DENSIDADE

DBSCAN - Como funciona?



# ANÁLISE DE AGRUPAMENTOS

## MÉTODO POR DENSIDADE

**dbscan**: *Density-based Clustering with R*

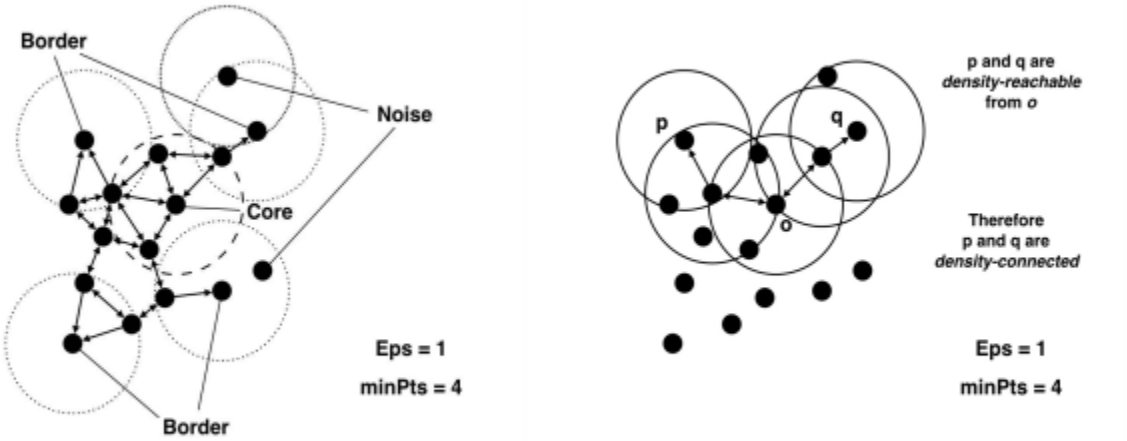



Figure 1: Concepts used the DBSCAN family of algorithms. (a) shows examples for the three point classes, core, border, and noise points, (b) illustrates the concept of density-reachability and density-connectivity.

Fonte: http: //cran.r-project.org/web/packages/dbscan/vignettes/dbscan.pdf

# ANÁLISE DE AGRUPAMENTOS

## Comparação entre as técnicas de CLUSTER

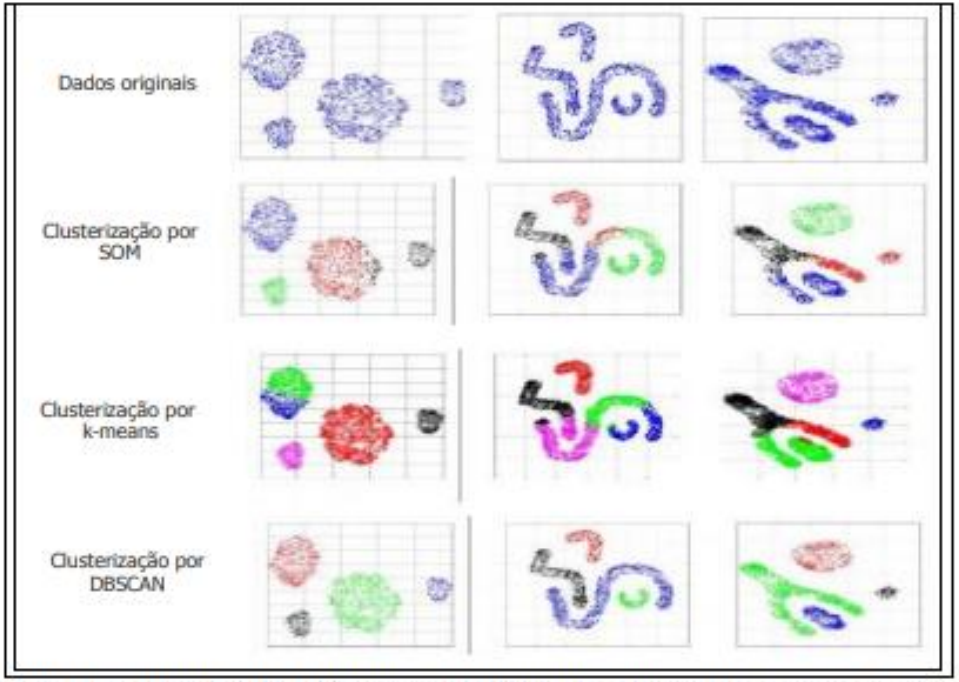


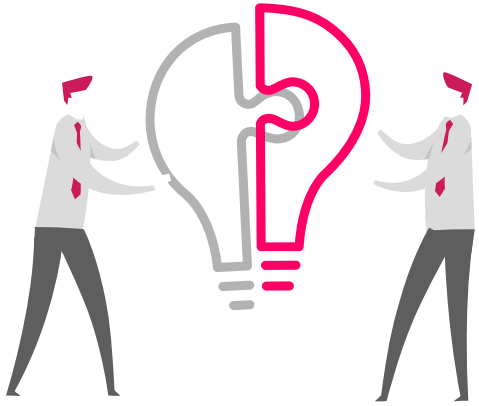

**Figura 5.9:** Desempenho de Diferentes Métodos de Clusterização para Dados Espaciais

Fonte: MUNTAZ & DURAISWAMY (2010).

Fonte: http://www.maxwell.vrac.puc-rio.br/24787/24787_6.PDF

# EXERCITANDO

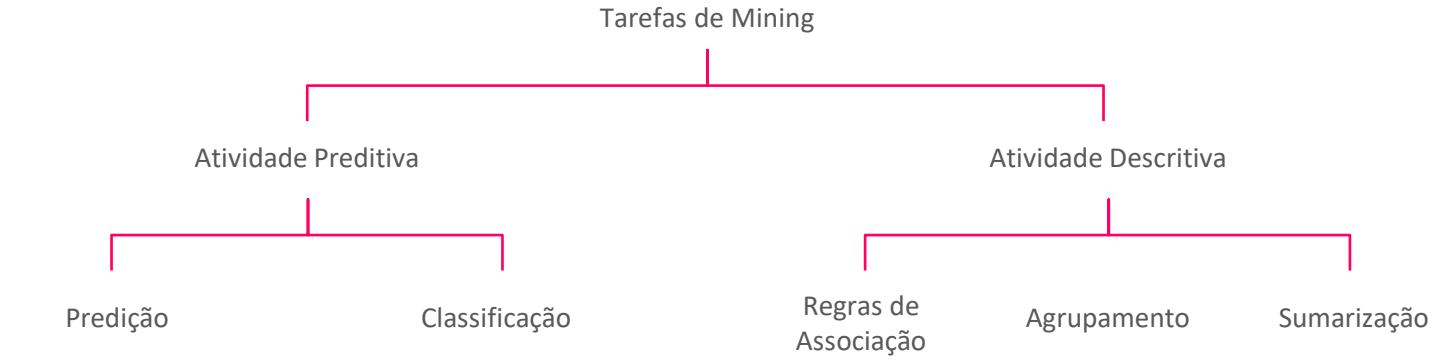
Cluster Analysis



Base
Países

# DATA MINING

Aplicações práticas de Data Mining podem ser categorizadas de acordo com a tarefa que se pretende resolver.



# ANÁLISE NÃO SUPERVISIONADA

## REDUÇÃO DA DIMENSIONALIDADE

### COMPONENTES PRINCIPAIS

### PCA

# Como definir um indicador de valor para os  clientes da empresa ACME?

# ANÁLISE DE COMPONENTES PRINCIPAIS PCA

- Técnica de aprendizado não supervisionado.
- O objetivo é encontrar combinações lineares das variáveis que incluam a maior quantidade possível de variância original das variáveis.
- Esta transformação é definida de forma que o primeiro componente principal tem a maior variância possível, e cada componente seguinte, por sua vez, tem a máxima variância sob a restrição de ser ortogonal a (i.e., não correlacionado com) os componentes anteriores.

Quanto maior a dimensão dos dados (número de variáveis) maior o risco de sobre ajuste do modelo.

Uma das razões pela qual a ACP é tão utilizada, é o fato obter componentes principais não correlacionadas. (alguns algoritmos conseguem melhor performance preditiva com variáveis com baixa correlação) .

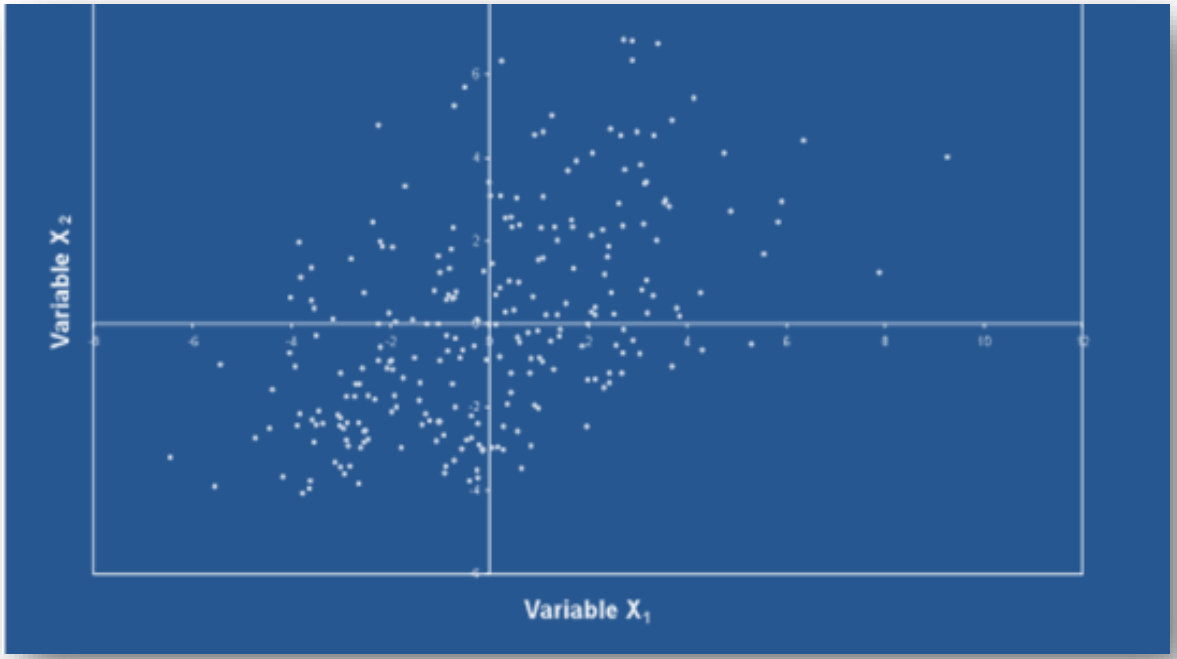
Outra forma de diminuir a presença de variáveis com alta colinearidade é excluí-las. Variáveis colineares trazem informação redundante(tempo perdido). Aumentam a instabilidade dos modelos.

# ANÁLISE DE COMPONENTES PRINCIPAIS PCA

- Transforma um conjunto de p variáveis originais em um novo conjunto de p variáveis, de variância máxima. Sendo assim, em geral é possível passar a trabalhar com um número bastante reduzido das novas variáveis, mantendo praticamente a mesma quantidade de informação.
- Objetivos gerais:
  - Redução dos dados
  - Interpretação
- É mais um *meio* do que um *fim*. Utilizada como passo intermediário antes do uso de outras técnicas estatísticas: Regressão Múltipla, Cluster, Análise Fatorial.
- Sempre que realizamos uma análise de componentes principais, esperamos conseguir explicar quase toda a variabilidade dos dados com uns poucos componentes principais.
- Pressupostos para aplicação da técnica : Variáveis em escala intervalar
- Sensível a diferenças de escala entre as variáveis
- Nem sempre é possível interpretar as componentes, atribuindo-se um nome

# ANÁLISE DE COMPONENTES PRINCIPAIS PCA

Exemplo:
Variável $X_1$ e $X_2$ tem covariância positiva e cada uma delas tem variância similar .
Cada variável é ajustada para ter média zero

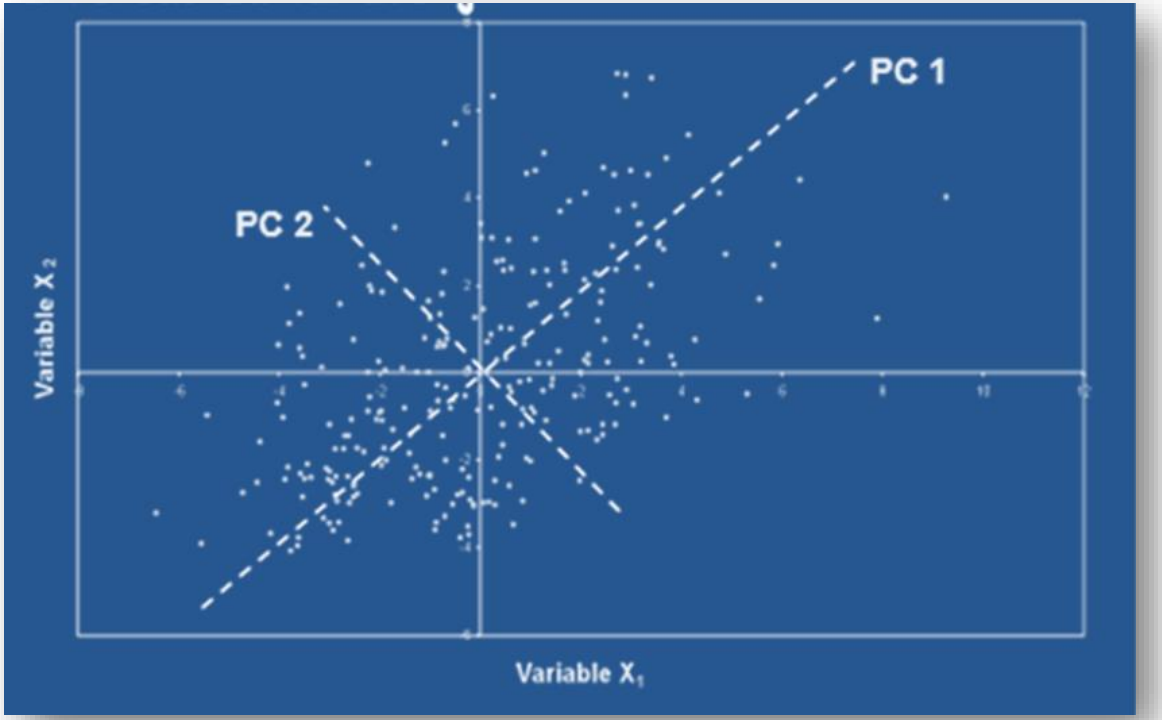


# ANÁLISE DE COMPONENTES PRINCIPAIS PCA

Cada eixo principal é uma combinação linear das variáveis originais

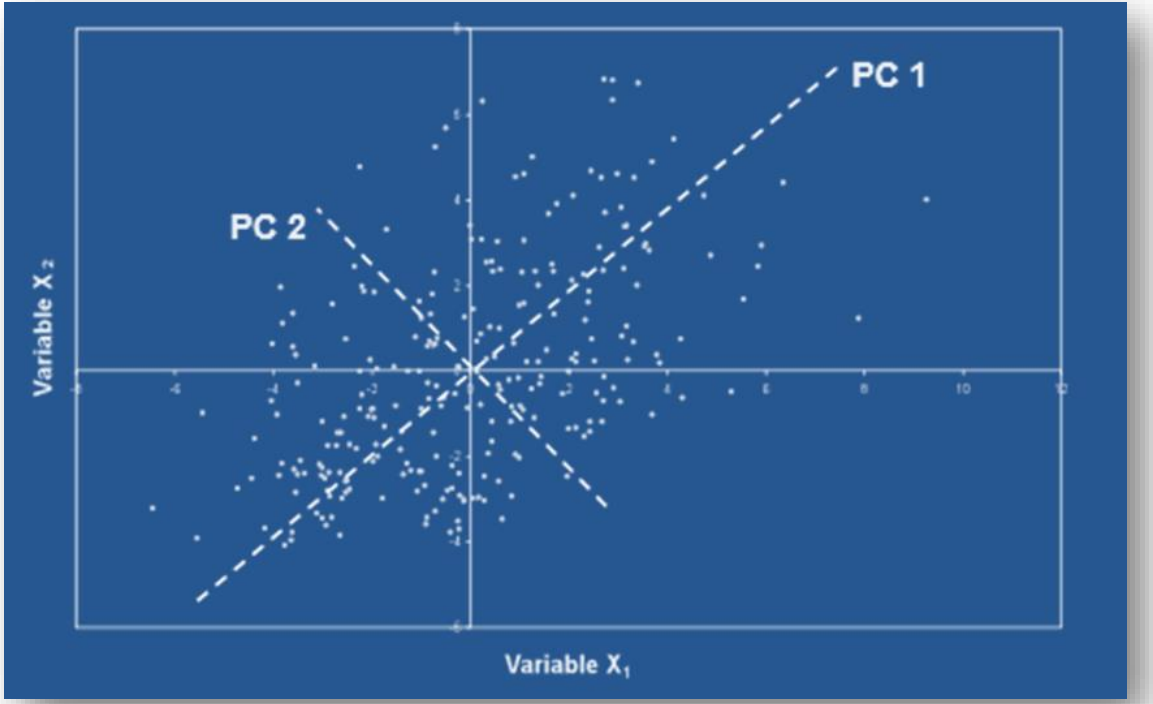
$PC_j = a_{i1}Y_1 + a_{12}Y_2 + ....a_{in}Y_n$

$a_{ij}$´s são os coeficientes para o fator i, multiplicado pela dimensão da variável j



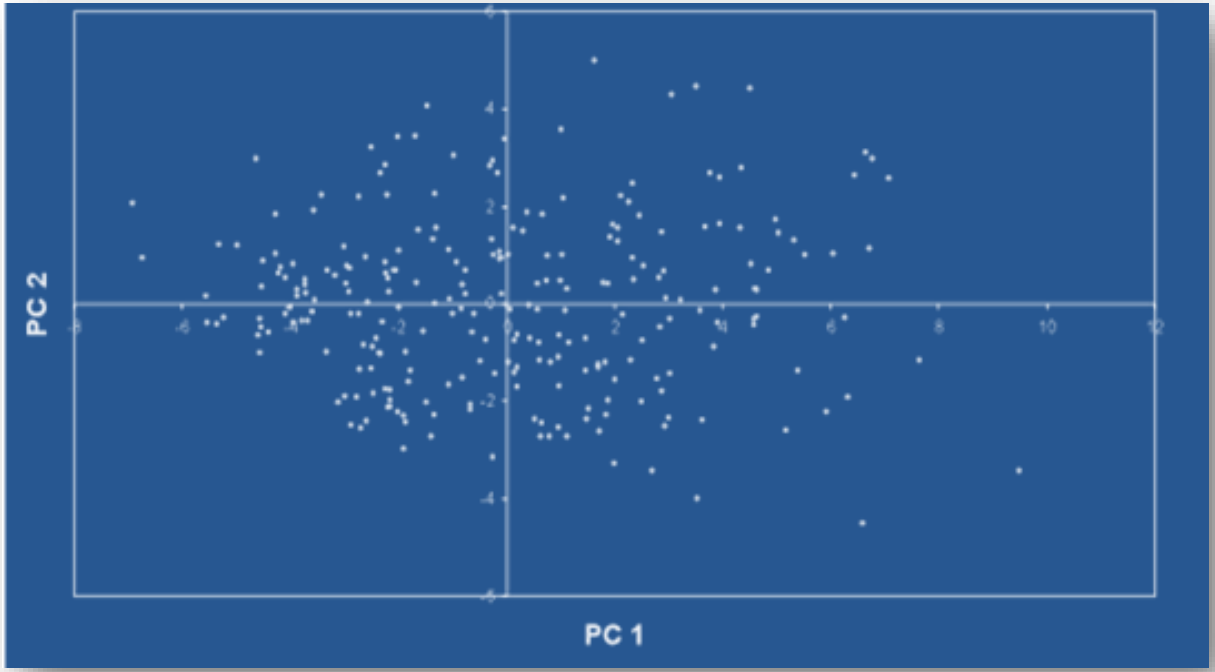
# ANÁLISE DE COMPONENTES PRINCIPAIS PCA

Os PC eixos são rotações rígidas das variáveis originais
PC1 é simultaneamente a direção de maior variância e simultaneamente melhor reta “ajustada” que minimiza a distância média entre os pontos e PC1



# ANÁLISE DE COMPONENTES PRINCIPAIS PCA

Componentes Principais são calculados
PC1 tem a maior variância possível
PC2 tem a segunda maior variância possível
PC1 e PC2 tem covariância zero.



# EXEMPLO: Projeto Fábrica da Cultura ( Fonte:Seade)

Desenvolvido pela Secretaria de Estado da Cultura, no município de São Paulo, foi criado o índice de vulnerabilidade juvenil (IVJ), cuja função central é auxiliar na escolha de áreas de intervenção, ou, no presente caso, os 96 distritos administrativos do município.

A escolha do termo *"vulnerabilidade juvenil" foi uma opção àqueles utilizados de forma mais recorrente, como "adolescentes em situação de risco" ou "adolescentes em situação de exclusão social", que, na ótica do projeto, poderiam distorcer o entendimento da grave e complexa problemática em que estão envolvidos os adolescentes.*

*A discussão da associação entre adolescência e "problemas/perigo", como decorrente tanto de fatores de natureza biológica como da autonomia relativa e ambígua que os jovens desfrutam na família e na sociedade, é uma preocupação presente nas literaturas médica e sociológica e na mídia. Da mesma forma, há o entendimento de que este fenômeno surge em sociedades modernas, acentuando-se em processos de rápida urbanização. Em outros termos, existe um vasto consenso de que a adolescência/juventude é um período de intensa vulnerabilidade.*

# EXEMPLO: Projeto Fábrica da Cultura ( Fonte:Seade)

A partir desta perspectiva, foi criado o índice de vulnerabilidade juvenil (IVJ), que considerou em sua composição os níveis de crescimento populacional e a presença de jovens entre a população distrital, freqüência à escola, gravidez e violência entre os jovens e adolescentes residentes no local. Este indicador varia em uma escala de 0 a 100 pontos, em que o zero representa o distrito com menor vulnerabilidade e 100 o de maior.

As variáveis selecionadas para compor o índice são:

- taxa anual de crescimento populacional entre 1991 e 2000;
- percentual de jovens, de 15 a 19 Anos, no total da população dos distritos;
- taxa de mortalidade por homicídio da população masculina de 15 a 19 anos;
- percentual de mães adolescentes, de 14 a 17 Anos, no total de nascidos vivos;
- valor do rendimento nominal médio mensal, das pessoas com rendimento, responsáveis pelos domicílios particulares permanentes;
- percentual de jovens de 15 a 17 anos que não frequentam a escola.

# EXEMPLO: Projeto Fábrica da Cultura ( Fonte:Seade)

O índice de vulnerabilidade juvenil foi obtido a partir de um modelo de análise fatorial. Esta técnica é frequentemente utilizada na resolução de problemas envolvendo um certo número de variáveis, em que se deseja a redução deste número com a finalidade de facilitar o entendimento analítico dos dados. Assim, a partir de uma análise da matriz de correlação das diversas variáveis, é possível obter indicadores sintéticos, que consistem numa combinação linear das variáveis originais que as sintetizam e explicam.

A aplicação deste modelo nos dados gerou um indicador sintético, que é a combinação linear das seis variáveis descritas anteriormente, explicando 74,2% da variabilidade total dos dados..

# EXEMPLO: Projeto Fábrica da Cultura ( Fonte:Seade)

| Distritos do Município de São Paulo | População Total | Participação da População Jovem de 15 a 19 Anos, no Total de Jovens do Município | População de Jovens de 15 a 19 Anos | Taxa de Aual de Crescimento Populacional | Taxa de Aual de Crescimento Populacional (Escala 0 a 100) | Participação dos Jovens de 15 a 19 Anos no Total da População dos Distritos (%) | Participação dos Jovens de 15 a 19 Anos no Total da População dos Distritos (Escala de 0 a 100) | Taxa de Mortalidade por Homicídio da População Masculina de 15 a 19 Anos (por 100.000 Hab.) | Taxa de Mortalidade por Homicídio da População Masculina de 15 a 19 Anos (Escala de 0 a 100) | Proporção de Mães Adolescentes de 14 a 17 Anos, no Total de Nascidos Vivos (%) | Proporção de Mães Adolescentes de 14 a 17 Anos, no Total de Nascidos Vivos (Escala de 0 a 100) | Rendimento Nominal Médio Mensal das Pessoas Responsáveis pelos Domicílios Particulares Permanentes (R$) | Proporção de Jovens de 15 a 17 anos que não Freqüentam à Escola %) | Densidade Demográfica (Hab./Km2) | Taxa de Fecundidade das Adolescentes de 14 a 17 Anos (por 1.000 mulheres) | Proporção de Jovens, de 18 a 19 Anos, que não Concluíram o Ensino Fundamental (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Água Rasa | 85.896 | 0,70 | 6.966 | - 1,11 | 17 | 8,11 | 38 | 38,5 | 7 | 4,32 | 29 | 1.503,34 | 14,89 | 12.449 | 20,49 | 25,23 |
| Alto de Pinheiros | 44.454 | 0,32 | 3.218 | - 1,37 | 16 | 7,24 | 19 | 43,3 | 8 | 2,33 | 11 | 4.809,46 | 8,89 | 5.773 | 8,44 | 21,95 |
| Anhanguera | 38.427 | 0,36 | 3.554 | 12,78 | 100 | 9,25 | 64 | 195,6 | 37 | 6,37 | 47 | 677,93 | 34,06 | 1.154 | 31,54 | 48,68 |
| Aricanduva | 94.813 | 0,89 | 8.884 | - 0,17 | 23 | 9,37 | 66 | 113,9 | 21 | 7,29 | 55 | 1.007,46 | 22,38 | 14.366 | 33,36 | 35 |
| Artur Alvim | 111.210 | 1,07 | 10.576 | - 0,69 | 20 | 9,51 | 69 | 199,3 | 38 | 7,88 | 60 | 875,02 | 18,27 | 16.850 | 41,96 | 32,72 |
| Barra Funda | 12.965 | 0,10 | 1.005 | - 2,31 | 10 | 7,75 | 30 | 134,1 | 25 | 7,58 | 57 | 2.364,04 | 19,17 | 2.315 | 43,48 | 33,73 |
| Bela Vista | 63.190 | 0,43 | 4.221 | - 1,41 | 16 | 6,68 | 7 | 135,4 | 25 | 5,86 | 42 | 2.435,70 | 19,26 | 24.304 | 38,62 | 35,8 |
| Belém | 39.622 | 0,32 | 3.213 | - 2,52 | 9 | 8,11 | 38 | 41,3 | 8 | 5,09 | 36 | 1.604,41 | 15,92 | 6.604 | 31,96 | 40,37 |
| Bom Retiro | 26.598 | 0,21 | 2.128 | - 3,40 | 4 | 8,00 | 36 | 295,0 | 56 | 5,31 | 38 | 1.358,39 | 23,86 | 6.650 | 38,41 | 39,86 |
| Brás | 25.158 | 0,20 | 2.018 | - 3,19 | 5 | 8,02 | 36 | 133,9 | 25 | 7,41 | 56 | 1.240,11 | 24,39 | 7.188 | 71,15 | 44,47 |
| Brasilândia | 247.328 | 2,56 | 25.425 | 2,32 | 38 | 10,28 | 86 | 354,6 | 67 | 8,57 | 66 | 666,13 | 32,75 | 11.778 | 57,67 | 50,87 |
| Butantã | 52.649 | 0,43 | 4.307 | - 1,06 | 18 | 8,18 | 40 | 93,8 | 18 | 3,62 | 23 | 2.584,46 | 11,22 | 4.212 | 20,59 | 19,9 |
| Cachoeirinha | 147.649 | 1,52 | 15.075 | 1,82 | 35 | 10,21 | 85 | 283,2 | 53 | 8,52 | 66 | 874,21 | 28,24 | 11.101 | 48,51 | 46,72 |
| Cambuci | 28.717 | 0,23 | 2.303 | - 2,83 | 7 | 8,02 | 36 | 29,5 | 6 | 5,12 | 36 | 1.604,97 | 15,79 | 7.363 | 30,34 | 32,51 |
| Campo Belo | 66.646 | 0,52 | 5.152 | - 1,73 | 14 | 7,73 | 30 | 105,3 | 20 | 3,49 | 22 | 3.800,67 | 13,24 | 7.573 | 16,61 | 22,55 |
| Campo Grande | 91.373 | 0,79 | 7.885 | 1,24 | 31 | 8,63 | 50 | 176,7 | 33 | 3,58 | 22 | 2.345,07 | 11,68 | 6.975 | 17,54 | 22,44 |
| Campo Limpo | 191.527 | 1,99 | 19.727 | 2,08 | 36 | 10,30 | 87 | 189,9 | 36 | 7,79 | 59 | 958,78 | 24,62 | 14.963 | 42,55 | 43,43 |
| Cangaiba | 137.442 | 1,32 | 13.112 | 2,02 | 36 | 9,54 | 70 | 145,4 | 27 | 7,78 | 59 | 948,16 | 22,73 | 8.590 | 38,26 | 37,31 |
| Capão Redondo | 240.793 | 2,59 | 25.741 | 2,49 | 39 | 10,69 | 95 | 298,3 | 56 | 7,96 | 61 | 711,37 | 27,81 | 17.705 | 42,99 | 47,19 |

# EXEMPLO: Projeto Fábrica da Cultura ( Fonte:Seade)

Resultado da Análise Fatorial

Foi realizada uma análise fatorial por componentes principais, que forneceu um escore fatorial. O modelo obtido explicou 74,2% da variabilidade total.

| Variáveis | | Cargas Fatoriais | Coeficientes Padronizados (1) |
|---|---|---|---|
| $X_1$ | Percentual de Mães Adolescentes, de 14 a 17 Anos, no Total de Nascidos Vivos | 0,933 | 0,182 |
| $X_2$ | Percentual de Jovens de 15 e 17 Anos que não Freqüentam a Escola | 0,914 | 0,177 |
| $X_3$ | Percentual de Jovens, de 15 a 19 Anos, no Total da População dos Distritos | 0,911 | 0,176 |
| $X_4$ | Taxa de Mortalidade por Homicídio da População Masculina de 15 a 19 Anos | 0,836 | 0,162 |
| $X_5$ | Valor do Rendimento Nominal Médio Mensal das Pessoas com Rendimento, Responsáveis pelos Domicílios Particulares Permanentes | -0,819 | 0,159 |
| $X_6$ | Taxa Anual de Crescimento Populacional | 0,741 | 0,143 |

(1) Coeficientes Padronizados : A soma dos coeficientes totaliza um.

# EXEMPLO: Projeto Fábrica da Cultura ( Fonte:Seade)

Equação para Determinação de um Fator

$$F_i = a_1 X_1 + a_2 X_2 + a_3 X_3 + .....+ a_p X_p$$

$a_i$ = peso ou coeficiente do fator
$F_i$ = Fator estimado , i=1, 2,...,p
p= número de variáveis

Construção do Índice de Vulnerabilidade Juvenil

$$IVJ = 0,182X_{1p} + 0,177X_{2p} + 0,176X_{3p} + 0,162X_{4p} + 0,158(100-X_{5p}) + 0,143X_{6p}$$

# EXEMPLO: Projeto Fábrica da Cultura ( Fonte:Seade)

A partir desta escala de pontos, foram gerados cinco grupos de vulnerabilidade juvenil:

**Grupo 1: até 21 pontos** - engloba os nove distritos menos vulneráveis do município de São Paulo: Jardim Paulista, Moema, Alto de Pinheiros, Itaim Bibi, Pinheiros, Consolação, Vila Mariana, Perdizes e Santo Amaro;

**Grupo 2: de 22 a 38 pontos** - engloba os 21 distritos que se classificam em segundo lugar entre os menos vulneráveis: Lapa, Campo Belo, Mooca, Tatuapé, Saúde, Santa Cecília, Santana, Butantã, Morumbi, Liberdade, Bela Vista, Cambuci, Belém, Água Rasa, Vila Leopoldina, Tucuruvi, Vila Guilherme, Campo Grande, Pari, Carrão e Barra Funda;

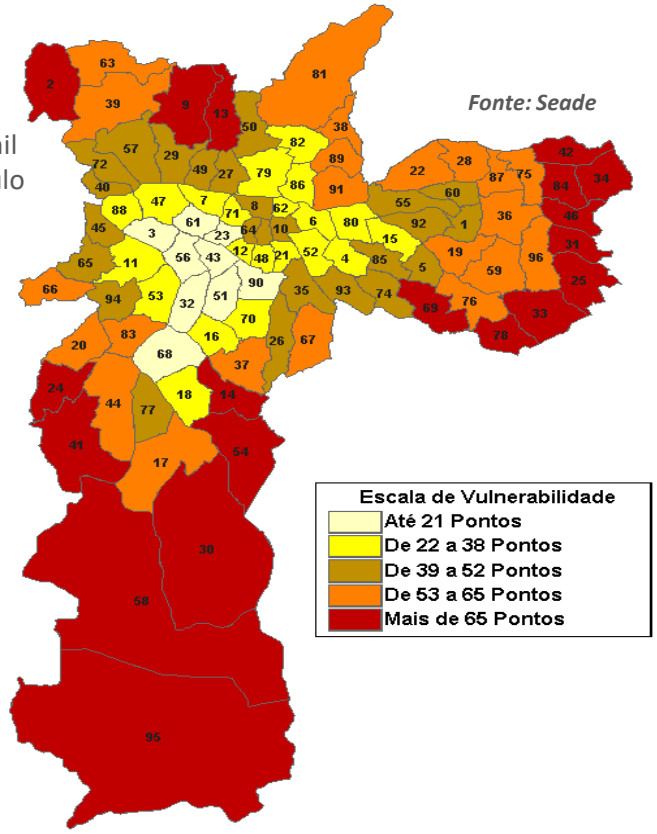
**Grupo 3: de 39 a 52 pontos** – engloba os 25 distritos que se posicionam em uma escala intermediária de vulnerabilidade: República, Penha, Mandaqui, Cursino, Socorro, Ipiranga, Casa Verde, Vila Matilde, Vila Formosa, Jaguara, Brás, Vila Prudente, Vila Sônia, Freguesia do Ó, Bom Retiro, São Lucas, Limão, São Domingos, Jaguaré, Rio Pequeno, Pirituba, Aricanduva, Sé, Artur Alvim e Ponte Rasa;

**Grupo 4: de 53 a 65 pontos** - engloba os 22 distritos que se classificam em segundo lugar entre os mais vulneráveis: Sacomã, Jabaquara, Vila Medeiros, Cangaíba, Cidade Líder, Vila Andrade, Vila Maria, Tremembé, Ermelino Matarazzo, São Miguel Paulista, José Bonifácio, Jaçanã, Itaquera, Raposo Tavares, Campo Limpo, São Mateus, Parque do Carmo, Vila Jacuí, Perus, Cidade Dutra, Jardim São Luís e Jaraguá;

**Grupo 5: mais de 65 pontos** - engloba os 19 distritos com maior vulnerabilidade juvenil do município de São Paulo: Cachoeirinha, Vila Curuçá, Guaianases, Sapopemba, Capão Redondo, Lajeado, Anhangüera, São Rafael, Jardim Helena, Cidade Ademar, Brasilândia, Itaim Paulista, Pedreira, Parelheiros, Jardim Ângela, Grajaú, Cidade Tiradentes, Iguatemi e Marsilac

# EXEMPLO: Projeto Fábrica da Cultura ( Fonte:Seade)

Grupos de Vulnerabilidade Juvenil
Distritos do Município de São Paulo



# BIBLIOGRAFIA


- KUHN, M. / JOHNSON K. Applied Predictive Modeling, 1st ed. 2013, Corr. 2nd printing 2018 Edition
- LESKOVEC, RAJAMARAM, ULLMAN. Mining of Massive Datasets, 2014. http://mmds.org.
- HAIR, J.F. / ANDERSON, R.E. / TATHAN, R.L. / BLACK, W.C. Análise multivariada de dados, 2009
- TORGO, L. Data Mining with R: Learning with Case Studies, 2.a ed. Chapman and Hall/CRC , 2007
- MINGOTI, S.A.; Análise de dados através de métodos de estatística multivariada, UFMG, 2005
- CARVALHO, L.A.V., Datamining – A mineração de dados no marketing, medicina, economia, engenharia e administração. Rio de Janeiro: Editora Ciência Moderna, 2005.
- BERRY,M.J.A., LINOFF,G. Data Mining Techniques For Marketing, Sales and Customer Support. 3a. ed. New York: John Wiley & Sons, Inc., 2011.
- DUNHAM, M.H. Data Mining - Introductory and Advanced Topics. Prentice Hall, 2002.
- DINIZ,C.A.R. , NETO F.L. Data Mining: Uma Introdução. São Paulo: XIV Simpósio Nacional de Probabilidade e Estatística. IME-USP, 2000.

# OBRIGADO

/AdelaideAlves

profadelaide.alves@fiap.com.br