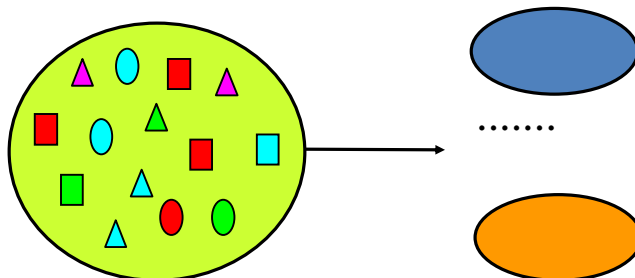


# MBA Big Data e Business Intelligence

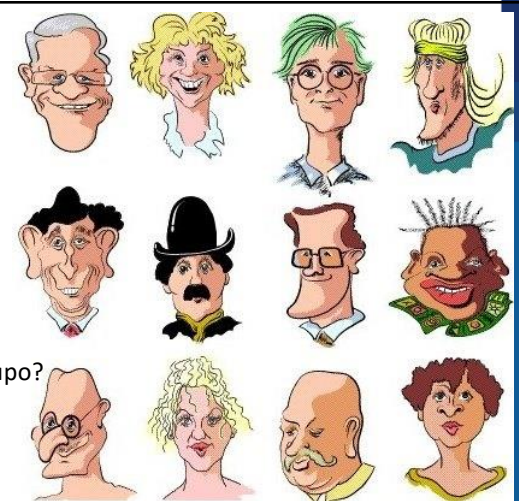
## Cluster Analysis

Prof. Abraham Laredo Sicsu

### Análise de agrupamentos (Cluster analysis)



3



**Agrupar essas pessoas:**

- ☐ Que critério utilizou?
- ☐ Quantos grupos formou?
- ☐ Quais os indivíduos de cada grupo?

1	2	3	4
5	6	7	8
9	10	11	12

3

4

## Discussão

FGV EDUCAÇÃO EXECUTIVA

- Quais as vantagens de agrupar os clientes de uma empresa em grupos homogêneos?
- Como / para que a empresa pode utilizar esses agrupamentos de clientes?

## Exercício

O diretor de RH deseja agrupar os funcionários da empresa em grupos homogêneos.

1. Como / para que o diretor de RH pode utilizar esses agrupamentos de funcionários?
2. Em que características dos funcionários deve basear-se para agrupá-los



## Aplicações de cluster analysis



- **Segmentação de mercados:**
  - Consumidores caracterizados por variáveis que expressam hábitos de consumo.
- **Classificação dos clientes de um banco**
  - Com base na distribuição de seus investimentos.
  - Com base nos serviços considerados importantes
- **Classificação de produtos:**
  - Os produtos de um mesmo grupo são percebidos como similares pelos consumidores potenciais.
- **Classificação de diferentes mercados (“praças”)**
  - Para analisar e definir estratégias mercadológicas.

## Aplicações de cluster analysis



### **Classificação de empresas**

com base em indicadores financeiros.

### **Classificação das perguntas de um questionário**

submetido a uma amostra piloto, para agrupar perguntas semelhantes e reduzir o questionário eliminando redundâncias.

### **Classificação dos funcionários de uma empresa**

a partir de variáveis que meçam seu relacionamento, envolvimento e fidelidade à mesma.

9

## Por que agrupar indivíduos ?

FGV EDUCAÇÃO EXECUTIVA

Agrupar indivíduos é uma necessidade básica em qualquer área de conhecimento.

- **Classificar** indivíduos de forma consistente.
- Síntese de **informação**:
  - A informação sobre N indivíduos é reduzida de forma conveniente à informação sobre apenas k grupos.
- “**Entender**” melhor a população em estudo.
- Elaborar / Confirmar **hipóteses**.
- Previsão do comportamento de **novos indivíduos**.

10

## Segmentação de mercado

FGV EDUCAÇÃO EXECUTIVA

**EXPORTAÇÕES EM \$mm**

**Agrupar países cujas exportações são similares**

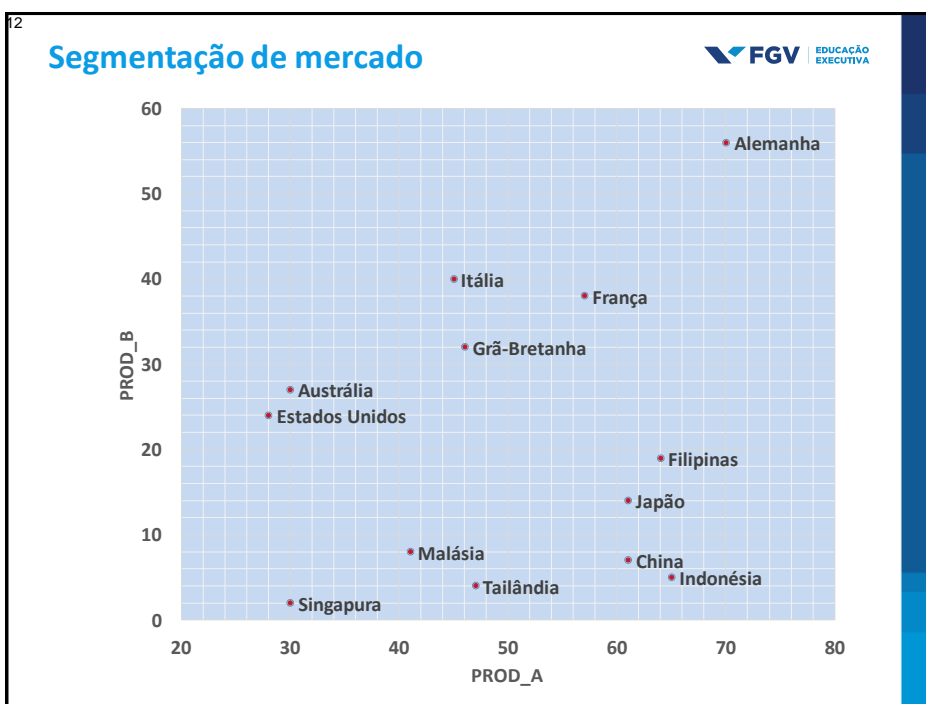
País	PROD_A	PROD_B
<i>Alemanha</i>	70	56
<i>Austrália</i>	30	27
<i>China</i>	61	7
<i>Estados Unidos</i>	28	24
<i>Filipinas</i>	64	19
<i>França</i>	57	38
<i>Grã-Bretanha</i>	46	32
<i>Indonésia</i>	65	5
<i>Itália</i>	45	40
<i>Japão</i>	61	14
<i>Malásia</i>	41	8
<i>Singapura</i>	30	2
<i>Tailândia</i>	47	4

11

## Segmentação de mercado

FGV EDUCAÇÃO EXECUTIVA

- Seus agrupamentos :



13

## Segmentação de mercado

País	PROD_A	PROD_B	PROD_C	PROD_D
Alemanha	43	56	122	37
Austrália	30	27	128	52
China	61	7	115	32
Estados Unidos	28	24	32	63
Filipinas	64	19	109	43
França	57	38	121	24
Grã-Bretanha	46	32	18	21
Indonésia	65	5	26	41
Itália	45	40	119	46
Japão	61	14	31	18
Malásia	41	5	5	56
Singapura	30	2	108	79
Tailândia	47	4	3	59

14

### Cluster Characteristics:



#### (2) Moderate Users:

- ❖ Purchase 6 - 30 times in last year.
- ❖ Purchase 4-9 different products.
- ❖ \$2,700 – next 5 years.

#### (3) Heavy Users:

- ❖ Purchase 31 or more times in last year.
- ❖ Purchase 10 or more different products.
- ❖ \$8,400 – next 5 years.

#### (1) Low /Non-Users:

- ❖ No or few purchases in last year (5 or <).
- ❖ Very low monetary value at present.

15

## Questões em Cluster Analysis



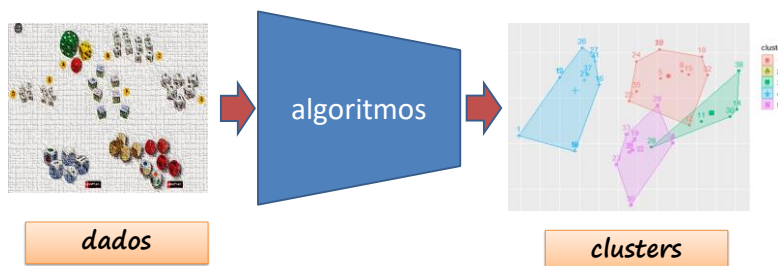
- Qual o **objetivo** do estudo?
- Que **variáveis** utilizar para caracterizar os indivíduos ?
- Como definir a **similaridade** entre os indivíduos ?
- Que **técnica** de agrupamento utilizar ?
- Em **quantos grupos** vamos agrupar os indivíduos ?
- Como **descrever** os grupos ?
- Como **validar** o resultado do agrupamento ?

16

## O grande dilema: clusters fazem sentido?



- Um algoritmo **sempre** gera clusters
- Mas ...será que por trás dos dados existe uma **estrutura de grupos**???
  - Os clusters **fazem sentido no contexto do problema**?????
  - Ou são meros resultados matemáticos????



*Na prática, nem sempre que fazemos uma análise de agrupamentos os resultados fazem sentido*



17

## Análise de agrupamentos - Roteiro



1. Definir objetivos do estudo.
2. Selecionar indivíduos a serem agrupados
3. Identificar variáveis (*drivers e discriminadoras*)
4. Coletar os dados
5. Analisar e tratar os dados
  - Outliers
  - Missing values
  - Transformação de variáveis
  - Correlações entre variáveis , etc.
6. Selecionar critério(s) de parença
7. Selecionar e aplicar algoritmo(s) de agrupamento
8. Identificar, analisar (interpretar) os agrupamentos
9. Validar resultados

18

## Exemplo de transformação 1



$X_1$  representada a aplicação em Poupança,  $X_2$  a aplicação em Renda Fixa e  $X_3$  a aplicação em Fundo de Ações. Valores em \$1000. Agrupar em função da distribuição dos investimentos.

**Qual o objetivo do banco?**

<i>Cliente</i>	$X_1$	$X_2$	$X_3$	<i>Total</i>
<i>A</i>	22	0	1	23
<i>B</i>	93	26	74	193
<i>C</i>	0	8	58	66
<i>D</i>	65	10	72	147
<i>E</i>	26	5	5	36
<i>F</i>	0	14	56	70
<i>G</i>	20	300	60	380
<i>H</i>	68	14	90	172
<i>I</i>	5	26	131	162
<i>J</i>	100	500	60	660
<i>K</i>	80	320	0	400
<i>L</i>	55	10	0	65

19

### Exemplo de transformação 1 (cont.)

Dados transformados considerando a % investida em cada aplicação.

<i>Cliente</i>	<i>Z<sub>1</sub></i>	<i>Z<sub>2</sub></i>	<i>Z<sub>3</sub></i>
<i>A</i>	96	0	4
<i>B</i>	48	13	38
<i>C</i>	0	12	88
<i>D</i>	44	7	49
<i>E</i>	72	14	14
<i>F</i>	0	20	80
<i>G</i>	5	79	16
<i>H</i>	40	8	52
<i>I</i>	3	16	81
<i>J</i>	17	76	9
<i>K</i>	20	80	0
<i>L</i>	85	15	0

20

### Medindo distâncias / proximidade



#### Dois desafios:

- 1) Medir a parecença entre os **indivíduos**
  - 1) Distância
  - 2) Similaridade
- 2) Medir a distância entre **clusters**

21

## Parecença de 2 indivíduos



$s_{ij}$  → semelhança entre indivíduo(i) e indivíduo (j)

$d_{ij}$  → distância entre indivíduo(i) e indivíduo (j)

Transformação  $d_{ij} = 1 - s_{ij}$  ou  $d_{ij} = 1 / s_{ij}$

### Variáveis quantitativas - exemplos

Distância euclidiana

Distância “city-block” (Manhattan)

### Variáveis qualitativas - exemplos

Coeficiente de concordâncias simples

Coeficiente de Jaccard etc.

22

## Parecença: Medidas de Distância



$X_1$  aplicação: poupança (R\$1000)

$X_2$  aplicação: dólares (US\$)

indivíduos	X1	X2
<b>A</b>	<b>150</b>	<b>1200</b>
<b>B</b>	<b>100</b>	<b>2000</b>
<b>C</b>	<b>100</b>	<b>1500</b>

### Distância euclidiana:

$$d(A,B) = \sqrt{(150 - 100)^2 + (1200 - 2000)^2} = 801.6$$

$$d(A,C) = \sqrt{(150 - 100)^2 + (1200 - 1500)^2} = 304.1$$

$$d(B,C) = \sqrt{(100 - 100)^2 + (2000 - 1500)^2} = 500.0$$

**Note que a distância é praticamente determinada por  $x_2$  !**

23

## Padronização de Variáveis Quantitativas



### alternativa 1

$$z_j = \frac{x - \bar{x}_j}{s_j}$$

Todas as variáveis  $Z_j$  ( $j=1,\dots,p$ ) terão mesma variância. Isto pode ser inconveniente.

Função **scale** do R

### alternativa 2

$$z_j = \frac{x_j - \min_j}{\max_j - \min_j}$$



Todos os valores ficam entre 0 e 1.

- Existem muitas formas de transformar os dados
- Transformação mais conveniente depende do problema em estudo

24

## Distância city-block (Manhattan)

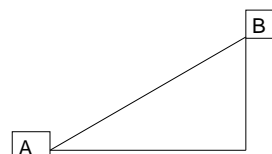


Não utiliza quadrados → reduz impacto de outliers

$$d(A, B) = \sum \|X_{Aj} - X_{Bj}\|$$

	X1	X2
A	150	12
B	100	20

$$d(A, B) = 50 + 8 = 58$$



25

## Parecença : Medidas de Similaridade



	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>	X <sub>5</sub>	X <sub>6</sub>	X <sub>7</sub>	X <sub>8</sub>
A	0	1	1	1	0	1	1	0
B	0	0	1	0	0	1	1	1

- Binárias **simétricas**: 0-0 e 1-1 tem mesma importância

- Podemos manter k - 1 dummies

$$\text{Concordâncias Simples} \Rightarrow S_1 = \frac{\text{concordâncias}}{p} = \frac{5}{8} = 0.63$$

- Binárias **assimétricas**: 0-0 e 1-1 não tem mesma importância

- Preferimos manter k dummies

$$\text{Concordâncias positivas} \Rightarrow S_2 = \frac{\text{concordâncias}(1-1)}{p} = \frac{3}{8} = 0.38$$

Concordâncias  
"importantes"

$$\text{Coeficiente de Jaccard} \Rightarrow S_3 = \frac{\text{concordâncias}(1-1)}{p - \text{concordâncias}(0-0)} = \frac{3}{6} = 0.50$$

26

## Similaridade entre qualitativas nominais



- S(Pedro, Maria) = 4/6

	E.Civil	Profissão	Residência	Residência	Sexo	Atividade
Pedro	Casado	Médico	SP	Alugada	Masc	Professor
Maria	Casada	Advogada	SP	Alugada	Fem	Professora

- Alternativa: transformar variável em dummies

## Similaridade entre qualitativas ordinais



- Variável ordenada em K categorias
- Alternativa 1: trabalhar com se fosse quantitativa (valores 1,2,...,K)
  - Vantagem: simplicidade
  - Problemas: diferença entre categorias por não ser igual
    - Secundário-primário  $\neq$  superior - secundário
- Alternativa 2: gerar dummies
  - Problema: perda da estrutura de ordem pode ser fatal
- Alternativa 3: imputar valores
  - Primário= 1, secundário = 3, superior= 9, pós graduado = 16
  - Depende da “experiência” do analista  $\leftarrow$  discutível

28

## Transformação de Variáveis Qualitativas



Estado civil	EC1	EC2	EC3	EC4
<b>Solteiro</b>	1	0	0	0
<b>Casado</b>	0	1	0	0
<b>Viúvo</b>	0	0	1	0
<b>Separado</b>	0	0	0	1

Qual das duas transformações adotar ?

- Binárias **assimétricas**: 0-0 e 1-1 não tem mesma importância
  - Preferimos manter k dummies
- Binárias **simétricas**: 0-0 e 1-1 tem mesma importância
  - Podemos manter k - 1 dummies

29

## Criação de variáveis - Dummies



```
>install.packages("dummies")
>library(dummies)
```

```
#gerando várias de uma vez
>dd1=dummy.data.frame(nu)
>names(dd1)
```

PRIMEIRO salvar  
arquivo nu com  
as.data.frame

```
[1] "STATUSbom" "STATUSmau" "IDADE" "UNIFEDMG" "UNIFEDRJ" "UNIFEDSC" "UNIFEDSP"
[8] "RESIDALUG" "RESIDMV" "RESIDOUTR" "RESIDPROP" "TMPRSD" "FONE" "ECIVCAS"
[15] "ECIVCASAD" "ECIVDIVORC" "ECIVNI" "ECIVOUTROS" "ECIVSOLT" "ECIVVIUVO" "INSTRUPRIM"
[22] "INSTRUSEC" "INSTRUSUP" "INSTRUNA" "RNDTOT" "RST2" "RSTnao" "RSTsim"
[29] "AGE"
```

Alternativa para k-1 dummies: utilizar `model.matrix` e deletar coluna de "1"

- veremos exemplo adiante

30

## Mistura de Variáveis de diferentes tipos



VAR.1 : Saldo médio na conta corrente ( R\$ 1000 )

VAR.2 : Tempo de conta (anos completos)

VAR.3 : Utiliza o cartão empresarial ( 1:sim; 0:não)

VAR.4 : Porte da empresa ( A; B; C; ME )

Empresa	VAR.1	VAR.2	VAR.3	VAR.4
Alfatec	40	12	1	B
Betausa	12	3	0	C
Gama, Inc.	120	6	1	A
Delta & Manos	2	0	0	ME

*Problema: como mesclar essa variáveis ?*

81

## Mistura de variáveis



- ❑ Discretizar variável contínua em duas ou mais categorias ordinais (quantis, por exemplo) e depois transformar em dummies.
  - problema: perda de informação
- ❑ Criar dois clusters (um só com as quanti e outro só com as quali) e cruzar as respostas.
  - Complexo!
- ❑ Qualitativas → dummies e quantitativas → padronizar entre 0 e 1 (ou 0-2)
- ❑ **Métrica de Gower (pacote "cluster" do R)**
  - combina os diferentes tipos de parença
  - veremos exemplo adiante

82

## Matriz de Similaridades &



## Matriz de Distâncias

Exemplo : Matriz de Distâncias entre países (baseada em indicadores econômicos)

	C.Rica	Brasil	Austr.	URSS	Urug.	A. Saud.	Japão	Niger
C. Rica	0							
Brasil	1.08	0						
Austrália	1.90	2.16	0					
URSS	0.74	0.93	1.48	0				
Uruguai	0.49	0.82	1.66	0.27				
A. Saudita	1.95	1.21	2.01	1.65	1.66	0		
Japão	2.23	2.51	0.37	1.84	2.02	2.27	0	
Niger	3.93	2.98	4.86	3.89	3.79	2.92	5.15	0



33

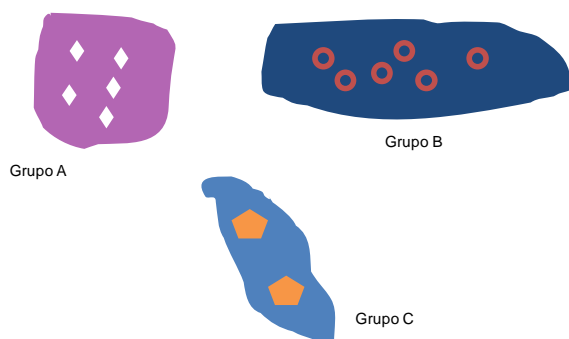
## Fusão de clusters - regras de ligação

- Como seleccionar clusters que devem ser agrupados (“ligados”)?

- A e B?

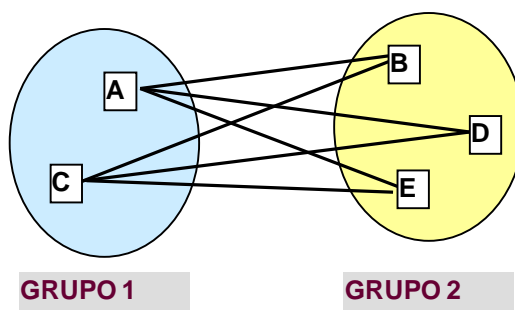
- A e C?

- B e C?



34

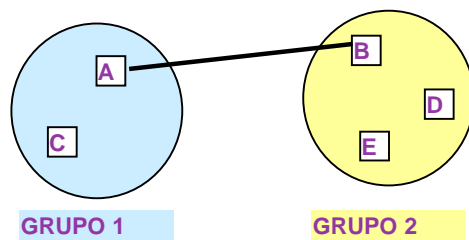
## Distância média



$$d(1,2) = \frac{d(AB) + d(AD) + d(AE) + d(CB) + d(CD) + d(CE)}{6}$$

35

### “Vizinho mais próximo” (VMP) (Single Linkage)

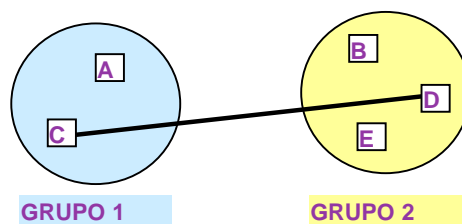


$$d(1,2) = \min [ d(AB), d(AD), d(AE), d(CB), d(CD), d(CE) ]$$

- Sensível a outliers
- Pode ocorrer efeito de encadeamento (chaining) afetando homogeneidade dos grupos. Ver referências.

36

### “Vizinho mais distante” (VMD) (complete linkage)



$$d(1,2) = \max [ d(AB), d(AD), d(AE), d(CB), d(CD), d(CE) ]$$

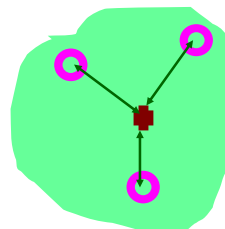
- Evita efeito de encadeamento
- Sensível a outliers

37

## SSE: Sum of Squared Errors (soma de quadrados residual)



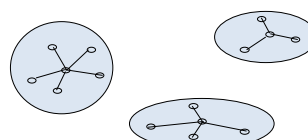
- Mede a “qualidade” de um agrupamento.
  - Mede a homogeneidade de cada agrupamento somando as dispersões dentro de cada cluster
  - Dados dois agrupamentos diferentes, preferimos o que tiver menor SSE



$C_i$ : cluster  $i$  ( $i=1, \dots, k$ )

$g_i$ : centroide de  $C_i$  (média dos elementos de  $C_i$ )

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} d(x, g_i)^2$$



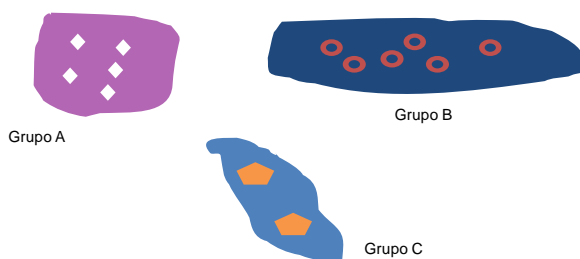
$d(x, g_i)$

38

## Método de Ward para agrupar dois clusters



- Calcular distância<sup>2</sup> de cada indivíduo ao centroide do grupo a que pertence.
- Calcular a soma de quadrados residual em cada cluster  $SSE_A$ ,  $SSE_B$ ,  $SSE_C$
- Calcular  $SSE_T : SSE_A + SSE_B + SSE_C$
- Fusão de grupos: a cada passo, fusão que provocar menor aumento na SSE



39

## Técnicas para Agrupar

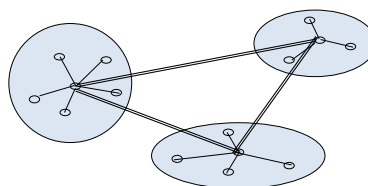
FGV EDUCAÇÃO EXECUTIVA

### .Coesão Interna

- Distância entre indivíduos de um grupo : pequena
- Homogeneidade “dentro” de cada agrupamento

### . Isolamento Entre Grupos

- Heterogeneidade entre agrupamentos
- Distância entre grupos : grande



Distâncias “dentro”  
Distâncias “entre”  
Centróides

40

## Classificação das técnicas deste curso

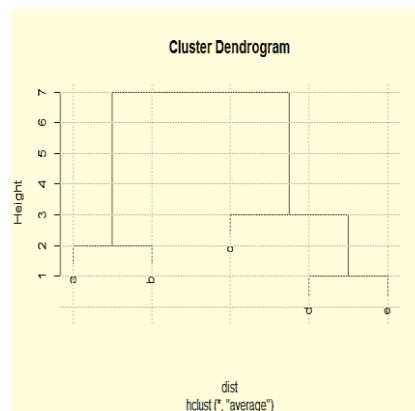
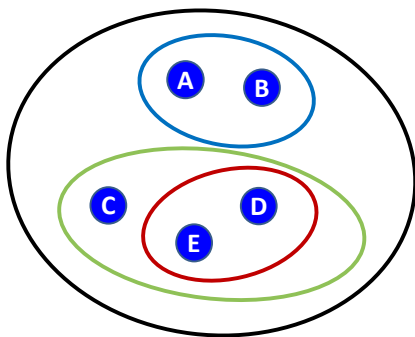
FGV EDUCAÇÃO EXECUTIVA

- Algoritmos hierárquicos aglomerativos
  - Ligação pela média (“average”, no R)
  - Método de Ward (Ward.D2 no R)
  - Outros
- Métodos de partição :
  - K-means
  - K-medoids
- DBSCAN (*density based...*)

41

## Algoritmos hierárquicos

- Passo 0: cada indivíduo é um cluster
- Passos seguintes: clusters vão sendo agrupados



42

## Agrupamento hierárquico com ligação pela média

**Matriz de  
distâncias**

$D_{ij}$	A	B	C	D	E
A	0.0				
B	2.0	0.0			
C	4.0	10.0	0.0		
D	6.0	8.0	2.0	0.0	
E	8.0	6.0	4.0	1.0	0.0

43

## Continuação

	<i>A</i>	<i>B</i>	<i>C</i>	<i>DE</i>
<i>A</i>	0.0			
<i>B</i>	2.0	0.0		
<i>C</i>	4.0	10.0	0.0	
<i>DE</i>	7.0	7.0	3.0	0.0



	<i>AB</i>	<i>C</i>	<i>DE</i>
<i>AB</i>	0.0		
<i>C</i>	7.0	0.0	
<i>DE</i>	7.0	3.0	0.0

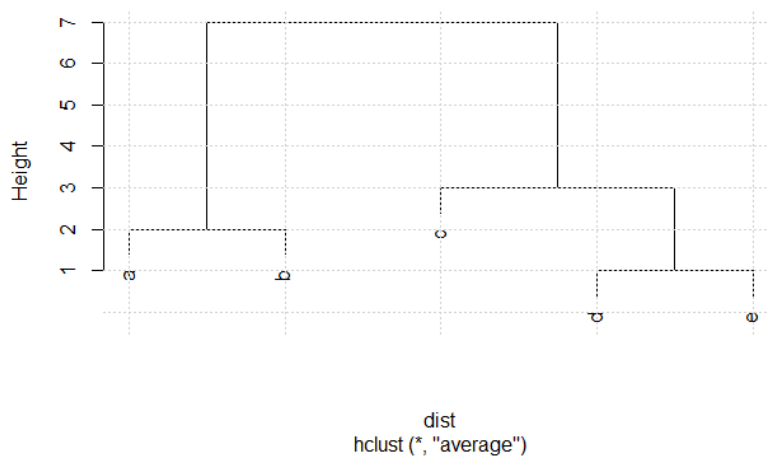


	<i>AB</i>	<i>CDE</i>
<i>AB</i>	0.0	
<i>CDE</i>	7.0	0.0

44

## Dendrograma

### Cluster Dendrogram



45

## Aplicação com variáveis quantitativas



- Dados: USArrests (biblioteca do R) + outras informações

crimes	arrests per 100,000 residents
Urbanpop	percent of the population living in urban areas
SFR	Standard federal regions (adicionado pelo Professor)

state	murder	assault	rape	UrbanPop	SFR
Alabama	13,2	236	21,2	58	IV
Alaska	10	263	44,5	48	X
Arizona	8,1	294	31	80	IX
Arkansas	8,8	190	19,5	50	VI
.....					

DRIVERS

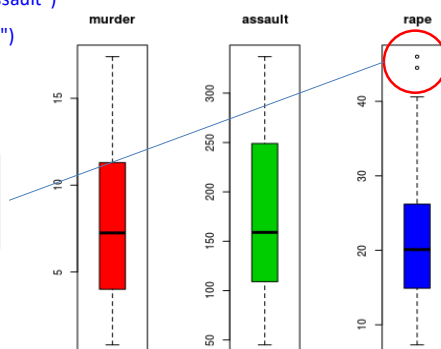


## USArrests



```
> us=USArrests[,1:6] # para facilitar digitação
> usdrivers=us[,2:4]
> par(mfrow=c(1,3))
> boxplot(usdrivers$murder, col=2, main="murder")
> boxplot(usdrivers$assault, col=3, main="assault")
> boxplot(usdrivers$rape, col=4, main="rape")
```

*Distribuição assimétrica.  
Vamos manter os dois  
pontos fora do boxplot*



## USArrests

Matriz de correlações

```
> print(cor(usdrivers), digits = 2)
```

	murder	assault	rape
murder	1.00	0.80	0.56
assault	0.80	1.00	0.67
rape	0.56	0.67	1.00

- Se observarmos  $|r| > 0,9$  removemos uma das variáveis

## USArrest

Padronizando as variáveis

```
> us.scale=scale(usdrivers)
```

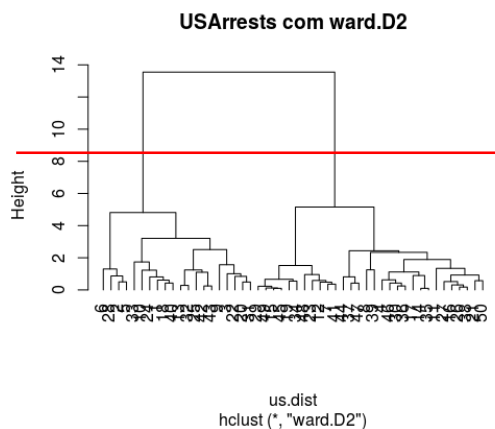
```
> head(us.scale)
```

	murder	assault	rape
[1,]	1.24256408	0.7828393	-0.003416473
[2,]	0.50786248	1.1068225	2.484202941
[3,]	0.07163341	1.4788032	1.042878388
[4,]	0.23234938	0.2308680	-0.184916602
[5,]	0.27826823	1.2628144	2.067820292
[6,]	0.02571456	0.3988593	1.864967207



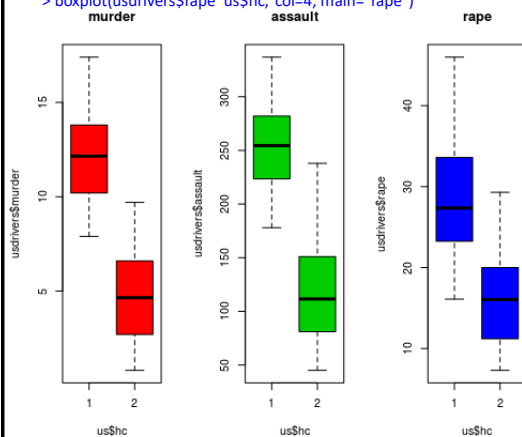
## USArrest

- Algoritmo hierárquico como método de Ward
- > `us.dist=dist(us.scale)` # matriz de distâncias
- > `hc=hclust(us.dist, method = "ward.D2")`
- > `plot(hc, hang=-1, main="USArrests com ward.D2")`



## USArrests

- > `us$hc=cutree(hc,2)` # selecionamos 2 clusters
- #Análise dos clusters com as drivers
- > `par(mfrow=c(1,3))`
- > `boxplot(usdrivers$murder~us$hc, col=2, main="murder")`
- > `boxplot(usdrivers$assault~us$hc, col=3, main="assault")`
- > `boxplot(usdrivers$rape~us$hc, col=4, main="rape")`



*Como descrever os clusters?*

## USArrests

```
> aggregate(us[,2:4],list(us$hc), median)
```

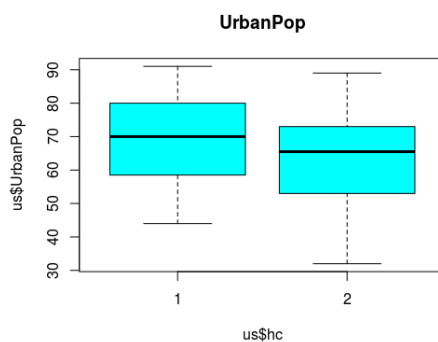
```
Group.1 murder assault rape
1      1  12.15   254.5  27.35
2      2   4.65   111.5  16.05
```

#Poderíamos tentar também max e min

## USArrests

- Análise dos clusters com as variáveis descritivas

```
> boxplot(us$UrbanPop ~ us$hc, col=5, main="UrbanPop")
```



## USArrets

```
> library(gmodels)
```

```
> CrossTable(us$SFR,us$hc, prop.c = F, prop.chisq = F, prop.t = F)
```

us\$SFR	us\$hc	1	2	Row Total
I	0	6	6	0.120
	0.000	1.000		
II	1	1	2	0.040
	0.500	0.500		
III	1	4	5	0.100
	0.200	0.800		
IV	7	2	9	0.180
	0.778	0.222		
IX	3	1	4	0.080
	0.750	0.250		
V	2	4	6	0.120
	0.333	0.667		
VI	3	2	5	0.100
	0.600	0.400		
VII	1	3	4	0.080
	0.250	0.750		
VIII	1	4	5	0.100
	0.200	0.800		
X	1	3	4	0.080
	0.250	0.750		
Column Total		20	30	50

Cuidado: as linhas estão em ordem alfabética

Quais as regiões mais perigosas?

## USArrests

### #componentes dos clusters

```
> us$state[us$hc==1]
```

```
"Alabama" "Alaska" "Arizona" "California" "Colorado" "Florida"
"Georgia" "Illinois" "Louisiana" "Maryland" "Michigan" "Mississippi"
"Missouri" "Nevada" "New Mexico" "New York" "North Carolina" "South Carolina"
"Tennessee" "Texas"
```

```
> us$state[us$hc==2]
```

```
"Arkansas" "Connecticut" "Delaware" "Hawaii" "Idaho" "Indiana" "Iowa" "Kansas"
"Kentucky" "Maine" "Massachusetts" "Minnesota" "Montana" "Nebraska"
"New Hampshire" "New Jersey" "North Dakota" "Ohio" "Oklahoma" "Oregon"
"Pennsylvania" "Rhode Island" "South Dakota" "Utah" "Vermont" "Virginia"
"Washington" "West Virginia" "Wisconsin" "Wyoming"
```

55

## Técnicas de Partição

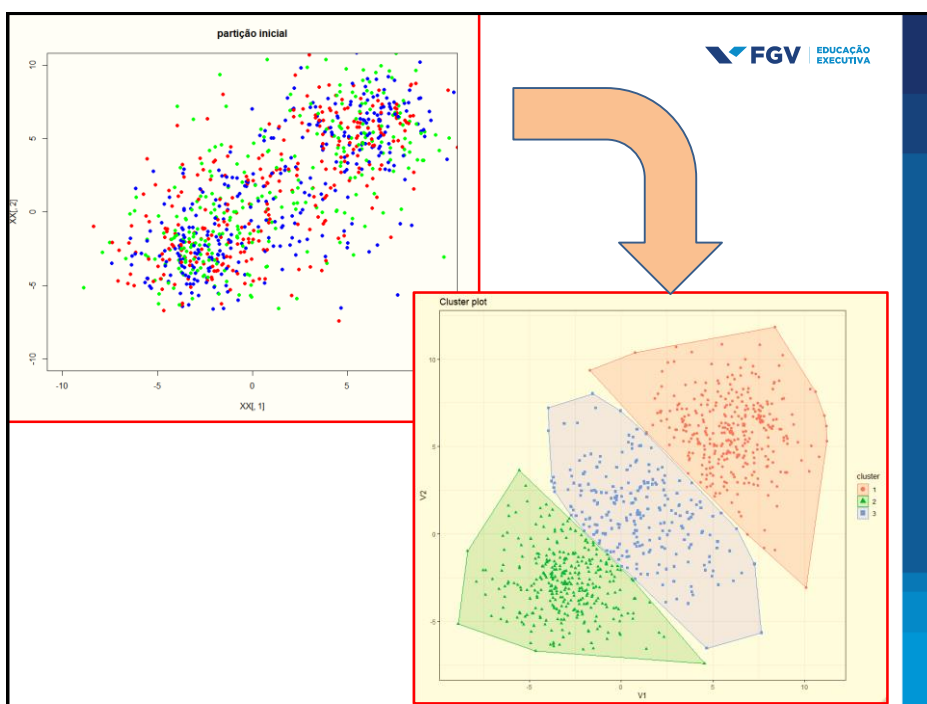
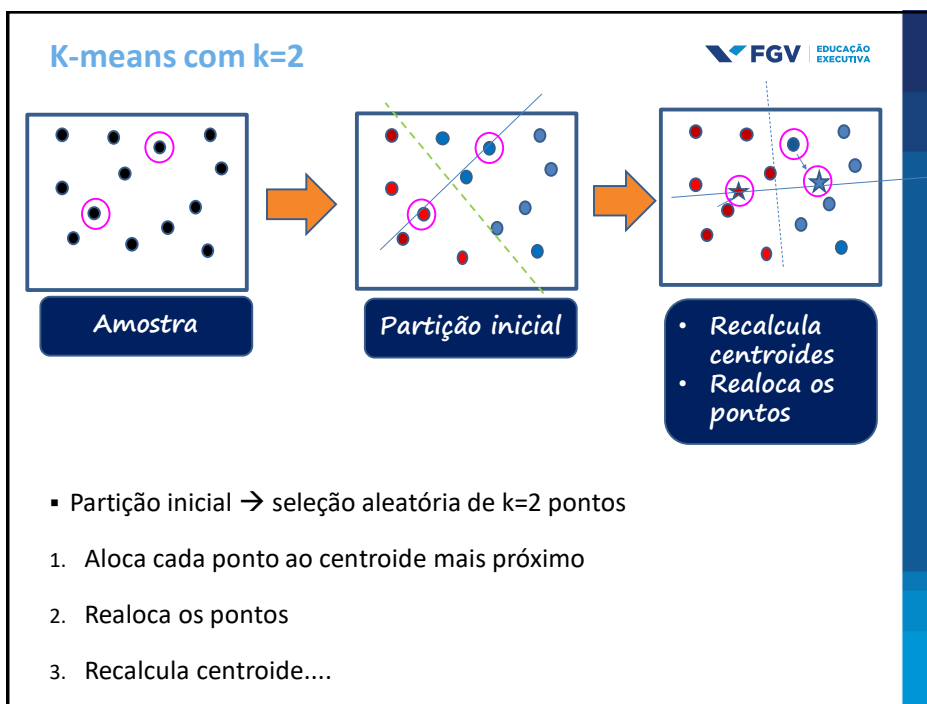


- Vamos nos limitar às técnicas *K-means* e *K-medoids*
- **Problemão:** Número de grupos (K) tem que ser fixado a priori.
- **Ideia:**
  - Partindo de uma *partição inicial*, **realocar** sucessivamente indivíduos entre grupos de acordo com objetivo pré - determinado. (ex.: minimizar a SSE).
- Determinar partição inicial
  - Várias formas
  - Solução é muito sensível á partição inicial
  - Podemos chegar a ótimos locais
  - Não é recomendado para detectar clusters com formatos não convexos

## K-means com k=2



- Partição inicial → seleção aleatória de k=2 pontos
1. Aloca cada ponto ao centroide mais próximo
  2. Realoca os pontos
  3. Recalcula centroide....



59

## número de grupos & partição inicial



### Determinação do número inicial de grupos K

- Várias formas
- R apresenta função `NbClust` com cerca de 30 critérios diferentes
- Experiência com o trabalho → intuição, perigo!
- Baseado na análise de dendrograma
- Testar para diferentes valores de K
  - Critérios quantitativos para avaliação e comparação das soluções
  - Avaliação subjetiva > Critério WOW

### Geração das partições iniciais - alternativas

- Default : seleção aleatória de k “centróides” ou Partição Aleatória
- Alternativas : ver leituras
- Solução é muito sensível á partição inicial

60

## Algoritmo básico do K-means

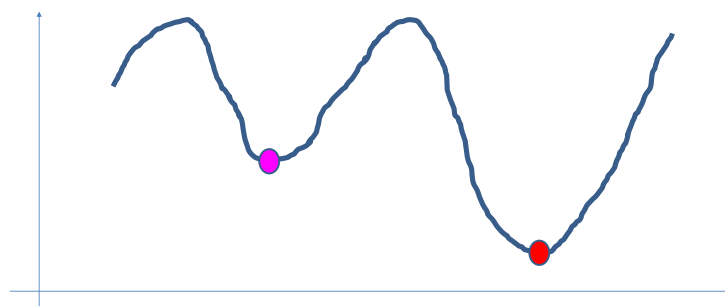


- **K-means: só deve ser utilizado com variáveis numéricas**
  - Depende do conceito de distância que não faz muito sentido no caso de dummies
- Passo 1: considerar centroides dos clusters iniciais
- Repita os passos seguintes
  - Passo 2: calcule as distâncias de cada indivíduo da amostra a cada um dos k centroides.
  - Passo 3: realoque cada indivíduo ao cluster de cujo centroide ele for mais próximo.
  - Passo 4: recalcule os centroides dos clusters obtidos após a realocação de todos os pontos
  - Passo 5: voltar ao Passo 2
  - Pare quando novos centroides forem os mesmos que os anteriores (ou quando o número máximo de iterações for atingido, ou quando redução da soma interna de quadrados for pouco significativa)
- **Algoritmo visa minimizar SSE, mas pode conduzir a ótimos locais**
  - Convém testar diferentes seleções iniciais de K sementes

51

## Ótimos locais

- Algoritmo visa minimizar SSE, mas pode conduzir a ótimos locais
  - Convém testar **diferentes partições iniciais**



Simplificação em 2 dimensões

## USArrests

```
> library(NbClust) # utiliza várias medidas da literatura para definir o número de clusters
```

```
> nb=NbClust(data=us.scale, diss= us.dist, distance = NULL, min.nc = 2,
max.nc = 8,method = "ward.D2", index = "all" )
```

Matriz distâncias

\*\*\*\*\*  
Among all indices:

- \* 14 proposed 2 as the best number of clusters
- \* 2 proposed 3 as the best number of clusters
- \* 4 proposed 4 as the best number of clusters
- \* 3 proposed 7 as the best number of clusters
- \* 1 proposed 8 as the best number of clusters

Resultado depende muito do max.nc fixado

Para kmeans com variáveis qualitativas não funciona

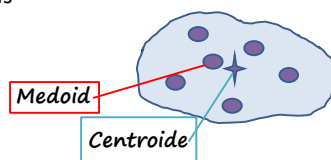
\*\*\*\*\* conclusion \*\*\*\*\*

\* According to the majority rule, the best number of clusters is 2

53

## k-medoids

- Similar ao K-means, mas cluster é representado por um de seus elementos e não pelo centroide
- Menos sensível à existência de outliers
- Medoid: **indivíduo do cluster** mais próximos dos demais (média das distâncias aos demais indivíduos)
- Cada cluster é representado por um de seus indivíduos (o medoid) e não pelo centroid
  - O medoid pode ser o indivíduo “típico” do cluster
- Adequado para trabalhar com mix de variáveis



54

## k-medoids

1. Selecionar k indivíduos como medoids (pode ser aleatoriamente)
  2. Alocar cada indivíduo da amostra ao medoid mais próximo
- Loop:
3. Em cada cluster determine o novo medoid
    3. Não houve alteração → pare. Fim do algoritmo
    4. Houve alteração de pelo menos um medoid → passo 4
  4. Realoque cada indivíduo da amostra ao cluster de cujo medoid for mais próximo.
  5. Volte à etapa 3



## USArrests



- Análise dos dados → já visto anteriormente
- Descrição dos clusters segue mesma linha do exercício anterior
- Daremos apenas os principais passos

## USArrests



#determinação do número de clusters com NbClust

#utiliza distância euclidiana, não precisamos utilizar opções diss e distance

```
> nk=NbClust(data=us.scale, min.nc = 2,max.nc = 8,method = "kmeans", index = "all")
```

\*\*\*\*\*

Among all indices:

- \* 14 proposed 2 as the best number of clusters
- \* 3 proposed 3 as the best number of clusters
- \* 1 proposed 5 as the best number of clusters
- \* 4 proposed 6 as the best number of clusters
- \* 2 proposed 8 as the best number of clusters

\*\*\*\*\* Conclusion \*\*\*\*\*

- \* According to the majority rule, the best number of clusters is 2

Mesmo número sugerido pelo dendrograma

## USArrests



#rodando com 2 clusters

```
> set.seed(18)
```

```
> kmn=kmeans(us.scale,2,nstart=25) # testa 25 partições iniciais
```

```
> us$kmn=kmn$cluster
```

```
> kmn$size
```

```
[1] 20 30
```

```
> kmn$centers
```

	murder	assault	rape
1	1.004934	1.0138274	0.8469650
2	-0.669956	-0.6758849	-0.5646433

## USArrests



#comparação k-means e hc

```
> table(us$hc,us$kmn)
```

	1	2
1	20	0
2	0	30

*Cuidado:*

- A numeração dos clusters pode diferir entre os dois métodos
- Nem sempre dá “certinho” como neste caso

#####

The adjusted Rand index is thus ensured to have a value close to 0.0 for random labeling independently of the number of clusters and samples and exactly 1.0 when the clusterings are identical (up to a permutation)

```
> library(fpc)
```

```
> clust_stats <- cluster.stats(diss, us$kmn, us$hc) # Corrected Rand index
```

```
> clust_stats$corrected.rand
```

Em nosso exemplo o coeficiente daria 1,0 (coincidência total)

69

## Caso CERVEJAS

### Segmentação com mix de variáveis



beer	rating	origin	price	cost	calories	sodium	alcohol	light
BUDWEISER	VeryGood	USA	2.59	0.43	144	15	4.70	NONLIGHT
SCHLITZ	VeryGood	USA	2.59	0.43	151	19	4.90	NONLIGHT
LOWENBRAU	VeryGood	USA	2.89	0.48	157	15	4.90	NONLIGHT
KRONENBOURG	VeryGood	France	4.39	0.73	170	7	5.20	NONLIGHT
HEINEKEN	VeryGood	Holland	4.59	0.77	152	11	5.00	NONLIGHT
OLD MILWAUKEE	Good	USA	1.69	0.28	145	23	4.60	NONLIGHT
AUGSBERGER	Good	USA	2.39	0.40	175	24	5.50	NONLIGHT
STROHS B. STYLE	Good	USA	2.49	0.42	149	27	4.70	NONLIGHT
MILLER LITE	Good	USA	2.55	0.43	99	10	4.30	LIGHT

drivers

Vamos exemplificar apenas k-medoids.

Algoritmo Hierárquico segue mesma linha anterior, a menos da geração da matriz de distâncias ← utilizar função daisy

Opção arbitrária, só para ilustrar procedimento

70

## Caso CERVEJAS



#vamos trabalhar com mistura de variáveis → library cluster necessária para calcular distância de Gower

```
> zz=CERVEJA
```

#transformando chr em factor para calcular distâncias com daisy

```
> zz$origin=as.factor(zz$origin)
```

```
> zz$light=as.factor(zz$light)
```

```
> cor(zz[,4:8])
```

script do R  
CERVEJAS.R

```

           price      cost  calories      sodium  alcohol
price  1.0000000  0.9998064  0.3301918 -0.4490626  0.3340570
cost    0.9998064  1.0000000  0.3238923 -0.4513526  0.3319127
calories 0.3301918  0.3238923  1.0000000  0.4124198  0.9209902
sodium  -0.4490626 -0.4513526  0.4124198  1.0000000  0.3214538
alcohol  0.3340570  0.3319127  0.9209902  0.3214538  1.0000000

```

71

## Caso MOBILE – mix de variáveis /77 países

FGV EDUCAÇÃO EXECUTIVA

GEOG	NIVDES	SISGOV	IDH	IALFAB	POP	IDADEMED	GROSSINC	MOBPHONE	INTERNET
Azerbaijan	Emerging	Parlamentarismo	0,70	1,00	9362,00	32,70	38187,80	1327,10	1955,00
China	Emerging	Outros	0,69	0,93	1354040,00	38,40	6321179,00	353892,50	193487,40
India	Emerging	Parlamentarismo	0,55	0,61	1245961,10	28,40	1659192,90	251090,90	24620,30
Indonesia	Emerging	Presidencialismo	0,62	0,92	247188,20	30,20	625885,40	52892,80	3457,30
Japan	Developed	Monarquia Constitucional	0,90	0,99	127342,50	45,00	4357995,90	40341,20	39641,70
Kazakhstan	Emerging	Parlamentarismo	0,71	1,00	16904,00	31,20	129184,60	3518,00	1958,40
Malaysia	Emerging	Monarquia Constitucional	0,74	0,92	29714,70	29,70	202851,30	8584,30	6083,80
.....									

- Drivers : considerar as variáveis
  - SISGOV **qualitativa**
  - IDH, IALFAB, IDADEMED, GROSSINC/POP, MOBPHONE/POP E INTERNET/POP **quantitativas**
- Como a variável NIVDES é função de outros indicadores quantitativos já considerados, será excluída
- Ver script **MOBIL.R**

72

## Utilização das Técnicas de Agrupamento

FGV EDUCAÇÃO EXECUTIVA

- Não há nenhuma técnica que seja sempre superior!**
  - Alguns estudos, tentando reproduzir estruturas de agrupamentos conhecidas, concluíram pela recomendação de K-means, Ward e ligação pela média, (Punj&Stewart-1983). **Não significa que são sempre melhores.**
  - K-means busca a melhor partição. Permite re-alocar elementos entre grupos. Métodos hierárquicos não permitem realocação.
- Recomendação:**
  - Rodar com diferentes técnicas e comparar resultados. Entender o porquê das inconsistências.
- Seleção da técnica tem maior influência no resultado que seleção do critério de parença (Punj & Stewart)
- Maior parte das técnicas é muito sensível a outliers. Dillon & Goldstein recomendam removê-los sempre. Discutir !

73

## Análise e validação - sugestões



- 1) Agrupar com diferentes distâncias e técnicas. Comparar resultados. Verificar consistência
- 2) Dividir amostra em duas partes. Rodar separadamente e comparar resultados. Identificar eventuais inconsistências.
- 3) Eliminar algumas variáveis arbitrariamente e comparar os diferentes resultados.
- 4) Alterar ordem dos indivíduos na matriz de dados para alterar seleção em casos de empates.
- 5) Existem indicadores e testes para verificar a consistência dos resultados. ( vide referências)

74

## Apêndice – distância de Gower



Da descrição de “daisy” no R

- Calcula a média das contribuições individuais  $d(ij,k)$  de cada variável  $k$  (onde  $i$  e  $j$  são duas observações)
- Cada variável quantitativa é padronizada entre 0 e 1 (subtraindo o mínimo e dividindo pela amplitude)
- A contribuição  $d(ij,k)$  de uma variável quantitativa é da diferença entre os valores dessa variável padronizada (manhattan!) entre as observações  $i$  e  $j$
- A contribuição de variáveis nominais ou binárias  $d(ij,k)$  é 0 se os dois valores forem diferentes e 1 se forem iguais
- As variáveis ordinais recebem um “valor inteiro” de 1 :  $m$  ( $m$ =categorias). Depois são tratadas como as quantitativas.
- A distância é a media das  $d(ij,k)$ 
  - Corresponde a ponderar por  $1/p$ .
  - Outros pesos podem ser atribuídos

## Apêndice – distância de Gower

### ▪ Matriz g

sex	salario	idade	fone	auto
s	1500	40	s	a
n	1400	50	s	b
s	1800	30	n	c

```
> g$sex=as.factor(g$sex)
> g$fone=as.factor(g$fone)
> g$auto=as.ordered(g$auto)
> dd2=daisy(g)
> as.matrix(dd2)
```

```
      1      2      3
1 0.00 0.45 0.65
2 0.45 0.00 0.90
3 0.65 0.90 0.00
```

76

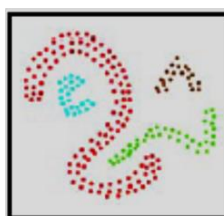
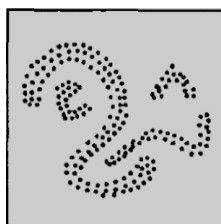
## MBA Big Data e Business Intelligence

### DBSCAN

Prof. Abraham Laredo Sicsu

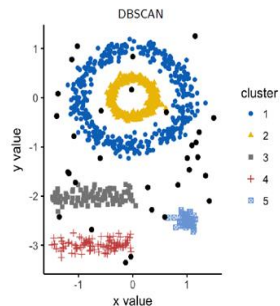
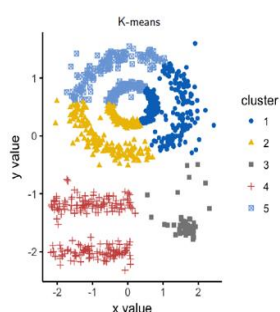
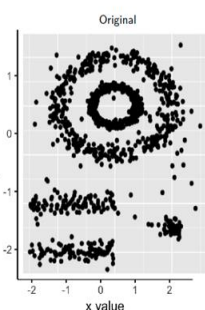
## Algoritmos Baseados em Densidade

- **Definição: Clusters baseados em densidade** são regiões de alta densidade de padrões separadas por regiões com baixa densidade, no espaço de padrões.
- Definição de densidade com base em agrupamento em torno de “centros”
- Segue ideia intuitiva do que seja um cluster



## Por que utilizar DBSCAN

- Métodos de partição (k-means, k-medoids,...) ou métodos hierárquicos funcionam bem quando os clusters são compactos e bem separados.
  - Não funcionam bem se formas dos clusters forem distintas dessas e na presença de outliers (que não deveriam ser incluídos em nenhum cluster)
- DBSCAN pode não ser a melhor solução!



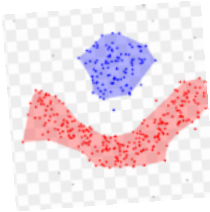
Pontos no data set “multishapes” do package factoextra do R

79

## Algoritmos Baseados em Densidade

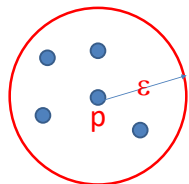
FGV EDUCAÇÃO EXECUTIVA

- **Definição: Clusters baseados em densidade** são regiões de alta densidade de padrões separadas por regiões com baixa densidade, no espaço de padrões.



**Parâmetros a serem definidos pelo analista**

- **$\epsilon$  (ou Eps)** : raio de uma região esférica (“vizinhança”) em torno do ponto **p**
- **$N_\epsilon(p)$** : quantidade de pontos que caem dentro da vizinhança de **p**, inclusive **p**



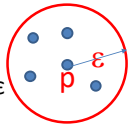
$N_\epsilon(p) = 5$

09/10/2020

## Definições

FGV EDUCAÇÃO EXECUTIVA

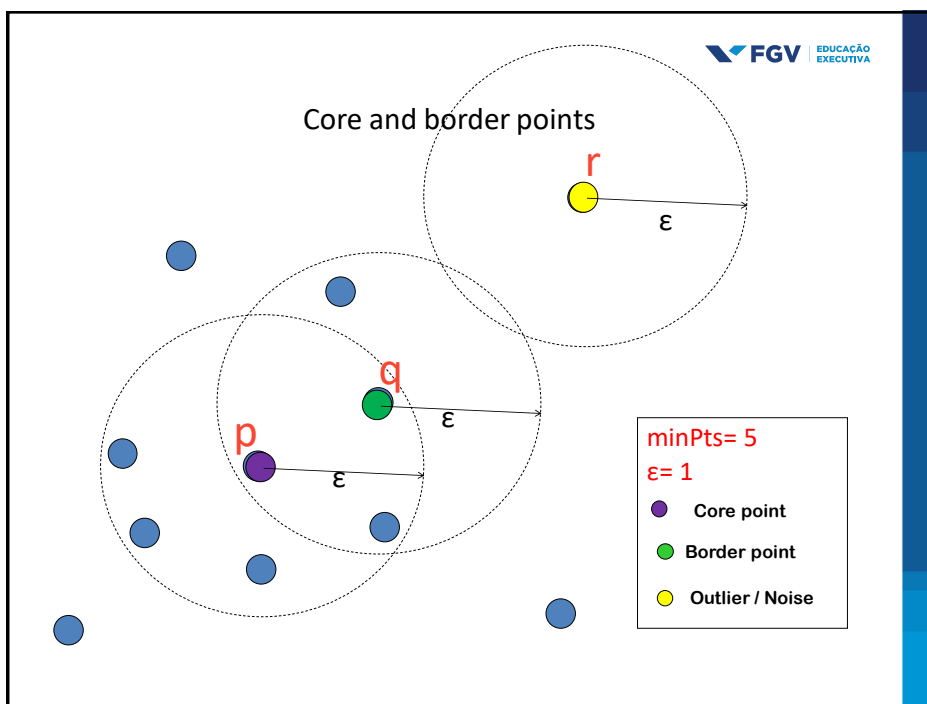
- **Definição : (  $\epsilon$ -vizinhança de um ponto)** A vizinhança de um objeto **p** com raio  $\epsilon$  é chamada de  $\epsilon$ -vizinhança de **p** é dada por:
  - $N_\epsilon(p)$  = quantidade de pontos que caem dentro da vizinhança de **p**, inclusive **p**
- **Definição : (Ponto Central / Núcleo – Core point)** : Se a  $\epsilon$ -vizinhança de um objeto **p** contém ao menos um número mínimo, *MinPts*, de objetos, então o objeto **p** é chamado de ponto central . (contagem inclui ponto **p**)
  - **p** é um ponto central sse  $N_\epsilon(p) \geq \text{MinPts}$
- **Definição : (Ponto de borda – Border point)**: Se a  $\epsilon$ -vizinhança de um objeto **p** contém menos que *MinPts* mas contém algum ponto central, então o objeto **p** é chamado de ponto de borda/ fronteira.
- **Definição : (outliers / noise)**: Se um ponto não for ponto central nem ponto de fronteira, epe será denominado outlier



$N_\epsilon(p) = 5$

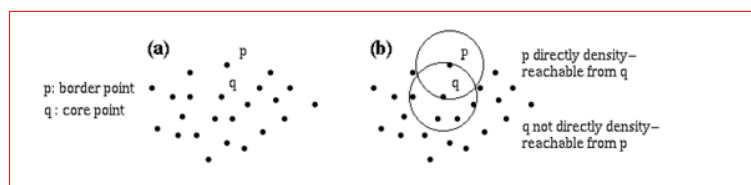
- Alguns textos utilizam a notação Eps em vez de  $\epsilon$





## Definição

- **Definição :** (Alcance Direto por Densidade / *directly density-reachable*): Um objeto **p** é alcançável por densidade diretamente do objeto **q**, com respeito à  $\epsilon$  e a *MinPts*, se:
  - **p** está na  $\epsilon$ - vizinhança de **q** e
  - **q** é um ponto central.
- O alcance direto por densidade não é simétrico se um ponto central e um ponto de borda estão envolvidos



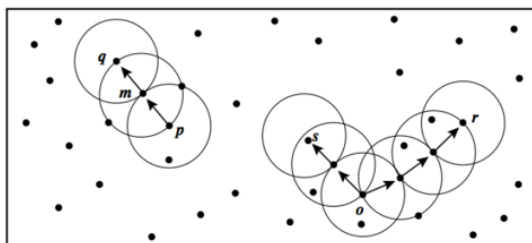


## Exercício

### Fonte:

Sarajane M. Pires e Clodoaldo A. M. Lima | Técnicas de Agrupamento (Clustering) | 17 de setembro de 2015 | 38 / 77

- Considere  $\text{MinPts}=3$ .
- Quais pontos são “centrais” (núcleos)
- Quais objetos são diretamente alcançáveis por densidade? Quais não são?
- Quais objetos são alcançáveis por densidade a partir de quais objetos? Quais não são?
- Quais objetos são conectados por densidade?



## Respostas

### DBSCAN

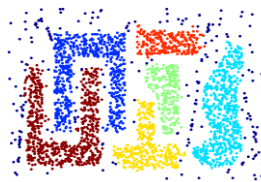
- **m, p, o e r** são objetos núcleos;
- **q** é diretamente alcançável por densidade a partir de **m**. **m** é diretamente alcançável por densidade a partir de **p** e vice-versa.
- **q** é (indiretamente) alcançável por densidade a partir de **p** porque **q** é diretamente alcançável por densidade a partir de **m** e **m** é diretamente alcançável por densidade a partir de **p**. Contudo, **p** não é diretamente alcançável por densidade a partir de **q** porque **q** não é um objeto núcleo. Similarmente, **r** e **s** são alcançáveis por densidade a partir de **o**, e **o** é alcançável por densidade a partir de **r**.
- **o, r, e s** são todos conectados por densidade.

## Clusters baseados em densidade - fundamentos

- Um cluster baseado em densidade é formado por um grupo de objetos conectados por densidade.
- Os algoritmos para agrupamentos baseados em densidade identificam regiões com alta densidade “cercadas” por regiões com baixa densidade
- Cada uma das regiões densas corresponde a um cluster



Pontos originais



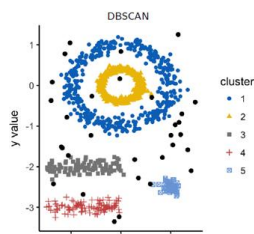
Clusters

## Definição de cluster density defined

**Definição 1:** um cluster com relação a  $\epsilon$  e MinPts é um subconjunto não vazio  $C$  da base de dados  $D$  para o qual valem as propriedades seguintes:

- 1) para todo par de pontos  $p$  e  $q$  de  $D$ : se  $p \in C$  e  $q$  for diretamente alcançável a partir de  $p$ , então  $q \in C$
- 2) para todo par de pontos  $p$  e  $q$  de  $C$ : o ponto  $p$  é conectado por densidade ao ponto  $q$

**Definição 2:** o conjunto de pontos de  $D$  que não pertencem a nenhum cluster  $C_1, \dots, C_k$  de  $D$  é denominado ruído (noise / outliers)



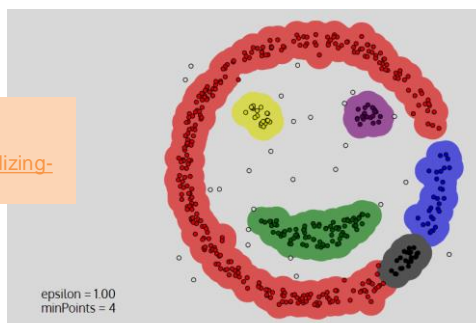
Pontos em preto: ruído

## Algoritmo DBSCAN – ideias

- DBSCAN: Density Based Spatial Clustering of Applications with Noise
- 2 pontos centrais (core points) próximos ( $d < \epsilon$ ) serão alocados ao mesmo cluster
- Border point de um core point será alocado ao mesmo cluster que o core point
  - Se border “pertence” a dois ou mais core points → regras de desempate
- Outliers são descartados
- Vai formando clusters “pulando de vizinho em vizinho”. Quando não houver mais vizinhos, começa outro cluster.

Ver “animação” do processo em

<https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/>



## Algoritmo DBSCAN – roteiro 1

- Roteiro do DBSCAN
  - Para cada ponto  $x$ , calcule a distância entre  $x$  e todos os demais pontos
  - Determine todos os vizinhos de  $x$  (pontos que caem dentro da vizinhança de  $x$  definida por  $\epsilon$ )
  - Se a vizinhança de  $x$  contiver um número de pontos maior ou igual a  $\text{MinPts}$ , então  $x$  é um core point
  - Para cada core point:
    - Se já pertence a um cluster, vá a outro ponto
    - Se não pertencer a um cluster previamente criado, crie um novo cluster incluindo esse ponto. Determine todos os pontos conectados por densidade a esse ponto e aloque-os ao mesmo cluster do core point.
  - Repita o procedimento até terminar de visitar todos os pontos.
  - Pontos que não pertencerem a nenhuma cluster são tratados como outliers
- Cada cluster consiste de todos os pontos conectados por densidade + pontos que estão em sua vizinhança (verificar???)

## Algoritmo DBSCAN – roteiro 2 (mais simples)



- Caracterizar cada um dos pontos como : central, fronteira, outlier
- Eliminar os outliers
- Unir progressivamente pontos centrais que distam entre si menos que  $\epsilon$
- Cada grupo de pontos centrais conectados forma um cluster
- Alocar cada ponto de fronteira ao cluster de um de seus correspondentes pontos centrais (em caso de empate → regras de desempate)

## Vantagens e desvantagens



- **Vantagens :**
  - Eficiente para agrupar grandes bases de dados
  - Permite obter clusters de formas diferentes
  - Adequado quando clusters não tem forma geométrica predefinida
  - Não requer especificação do número de clusters
  - Permite isolar outliers
- **Desvantagens :**
  - Muito sensível aos valores dos parâmetros  $\epsilon$  e **MinPts**
  - Pode produzir agrupamentos não confiáveis quando os clusters apresentam densidades significativamente diferentes

## DBSCAN: Sensitive to Parameters

Figure 8. DBSCAN results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

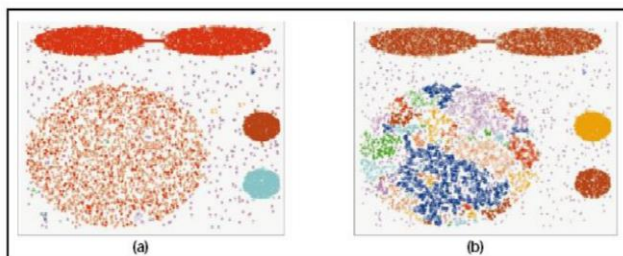
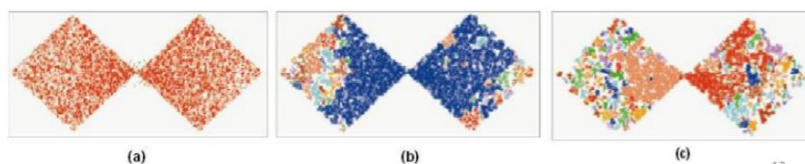
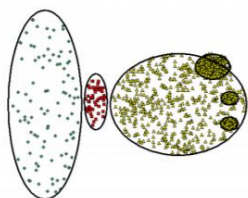


Figure 9. DBSCAN results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



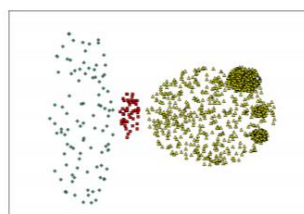
12

## When DBSCAN Does NOT Work Well

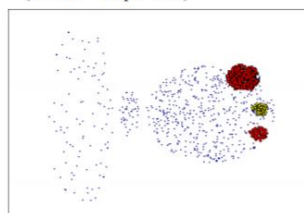


Original Points

- Cannot handle varying densities
- sensitive to parameters—hard to determine the correct set of parameters



(MinPts=4, Eps=9.92).



(MinPts=4, Eps=9.75)

15

## Exercício

- Fonte: <https://elvex.ugr.es/idbis/dm/slides/43%20Clustering%20-%20Density.pdf>

### Ejercicio

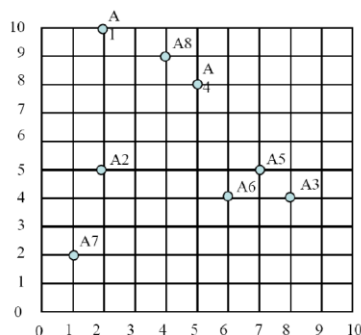
Agrupar los 8 puntos de la figura utilizando el algoritmo DBSCAN.

Número mínimo de puntos en el "vecindario":

$$\text{MinPts} = 2$$

Radio del "vecindario":

$$\text{Epsilon } \sqrt{2} > \sqrt{10}$$



## Ejercicio resuelto

Distancia euclídea

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

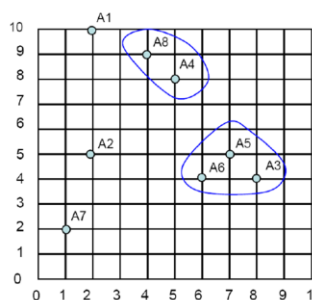




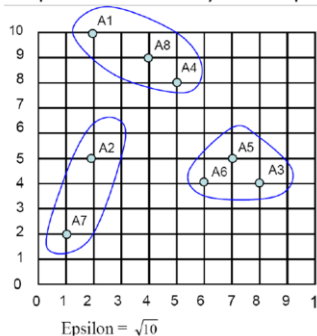
**Ejercicio resuelto**

Epsilon =

A1, A2 y A7 no tienen vecinos en su vecindario, por lo que se consideran "outliers" (no están en zonas densas):

**Ejercicio resuelto**Epsilon =  $\sqrt{10}$ 

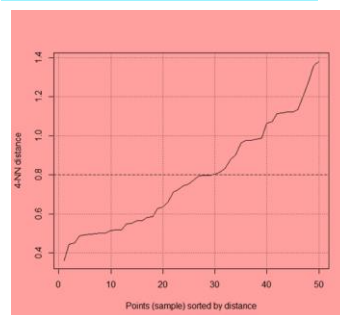
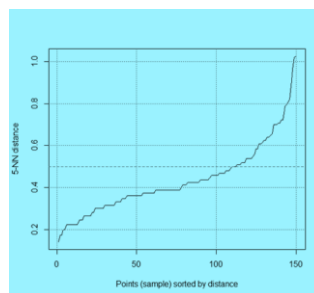
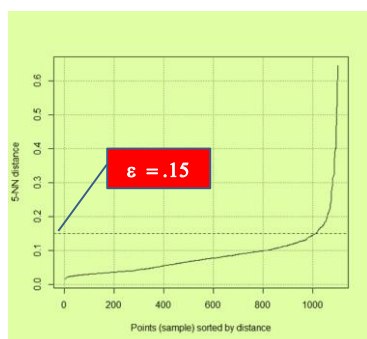
Al aumentar el valor del parámetro Epsilon, el vecindario de los puntos aumenta y todos quedan agrupados:



## Rodando DBSCAN no R

- Inicialmente selecionamos os parâmetros  $\epsilon$  e MinPts
- Cuidado: o resultado é  **muito**  influenciado pela escolha dos parâmetros
- Decisão pode ser por tentativa e erro , analisando os resultados obtidos
- **Regra prática sugerida em todos os textos:**
  - Fixe um valor k para minPts
  - Para cada ponto calcule k-dist, a distância até o k-ésimo vizinho mais próximo
  - Ordene e plote esses pontos.
  - Bons valores para  $\epsilon$  : onde a curva apresenta um cotovelo
    - Nem sempre é visível, ou há diferentes cotovelos
- Racional:
  - Para pontos de um mesmo cluster k-dist será pequeno se k não for maior que o tamanho do cluster.
  - Pontos fora do cluster, k-dist tende a ser grande
- Observação : variando k, o gráfico varia mas em geral a variação do  $\epsilon$  não é muito grande

## Exemplos de gráficos para determinar $\epsilon$



Gráficos  
obtidos  
diretamente  
com R

Corte difícil de  
decidir

## Rodando para arquivo íris (flores) no R



```
> data("iris")
> library(dbSCAN)
```

```
> db=iris[,1:4]
```

```
> db=scale(db)
```

#inicialmente vamos rodar com a matriz de dados

```
> db=as.matrix(db)
```

```
> kNNdistplot(db,k = 5)
```

```
> abline(h=.75, col = "red", lty=2)
```

```
> res <- dbSCAN(db, eps = .75, minPts = 5)
```

```
> res
```

DBSCAN clustering for 150 objects.

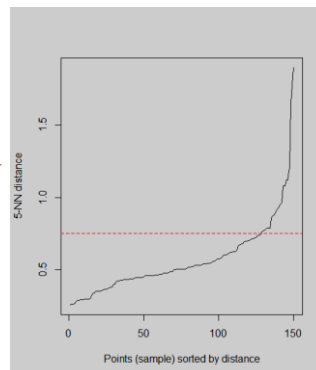
Parameters: eps = 0.75, minPts = 5

The clustering contains 2 cluster(s) and 5 noise points.

```
0 1 2
```

```
5 49 96
```

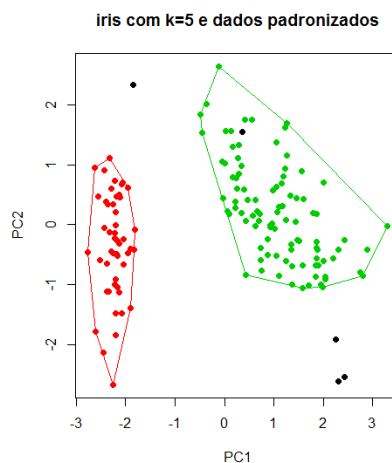
```
> cluster1=res$cluster
```



## Representação gráfica com pacote dbSCAN



```
> hullplot(db,res,lwd=1.5,cex=1.5,pch=20,solid = F, main = "iris com k=5 e dados padronizados")
```



## Rodando para arquivo íris (flores) no R- utilizando diretamente a matriz de distâncias



#calculando com matriz de distâncias ← vantagem: podemos utilizar com Gower

```
> diss=dist(db)
> kNNdistplot(diss, k = 5)
> abline(h=.75, col = "red", lty=2)
> res <- dbSCAN(diss, eps = .75, minPts = 5)
> res
```

DBSCAN clustering for 150 objects.  
Parameters: eps = 0.75, minPts = 5  
The clustering contains 2 cluster(s) and 5 noise points.

```
0 1 2
5 49 96
```

Available fields: cluster, eps, minPts

```
> dd=as.matrix(diss)
> cluster2=res.dis$cluster
> table(cluster1,cluster2)
```

```
      cluster2
cluster1 0  1  2
0         5  0  0
1         0 49  0
2         0  0 96
```



## Apêndice

Dados “multishape”

## Arquivo "multishapes" do R

```
> data("multishapes", package = "factoextra")
> plot(multishapes[, -3]) # não imprimir aqui
> db=multishapes[, -3]
> kNNdistplot(db, k = 5)
> grid(col=3)
> abline(h=.15, col = "red", lty=2)
> res <- dbSCAN(db, eps = .15, minPts = 5)
> res
```

DBSCAN clustering for 1100 objects.

Parameters:  $\text{eps} = 0.15$ ,  $\text{minPts} = 5$

The clustering contains 5 cluster(s) and 31 noise points.

```
0 1 2 3 4 5
31 410 405 104 99 51
```

```
> hullplot(db, res, solid = F, pch=19)
```

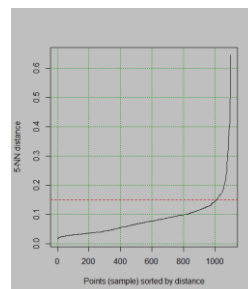
#utilizando os recursos de package factoextra

```
> fviz_cluster(res, data = db, ellipse = F, geom = "point",
+             show.clust.cent = F, palette="d3", ggtheme = theme_bw())
```

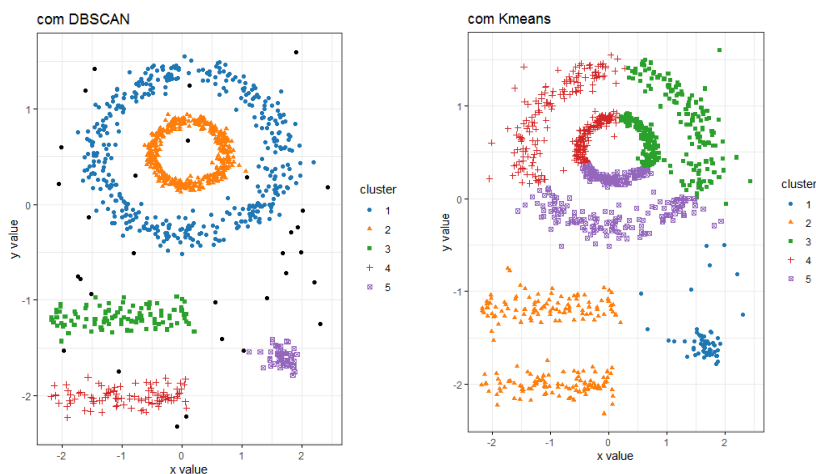
#com kmeans

```
> km=kmeans(db,5)
```

```
> fviz_cluster(km, data = db, ellipse = F, geom = "point",
+             show.clust.cent = F, palette="d3", ggtheme = theme_bw())
```



## Resultados



## Exercício



- Agrupar dados de **indicadoresdemograficos** com hclust, kmeans e dbscan
  - Utilizar package cluster para gerar matriz de distâncias
  - Fviz\_clust → utilizar matriz só com as variáveis numéricas
- Agrupar dados de **CERVEJAS** com hclust, kmeans e dbscan
  - Utilizar package cluster para gerar matriz de distâncias
  - Fviz\_clust → utilizar matriz só com as variáveis numéricas