

MAE 5776

# ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

[pavan@ime.usp.br](mailto:pavan@ime.usp.br)

1º Sem/2022 - IME

# Análise Multivariada

$$Y_{n \times p} = (Y_{ij}) \in \mathbb{R}^{n \times p}$$

Já vimos ☺

- ✓ Estatísticas descritivas multivariadas, Episódios de Concentração, Boxplot Bivariado
- ✓ Distribuição  $N_p$ , Distribuições Amostrais ( $T^2$  e  $W_p$ )
- Decomposições:  $SS_T$  e  $Y_{n \times p}$
- ✓  $N_p(\mu_g; \Sigma_g)$ : Inferências sobre  $\mu_g$  ( $T^2$ , MANOVA, ICS, Correções para Múltiplos testes)



## Técnicas Multivariadas:

Já vimos ☺

- ✓ 1. Análise de Componentes Principais (CP)
- ✓ 2. Escalonamento Multidimensional (CoP)
- ✓ 3. Análise de Correspondência
- ✓ 4. Análise Fatorial Exploratória (e Confirmatória)
- ✓ 5. Análise Discriminante: Solução de Fisher e Regra Geral de Bayes

- **Análise de Agrupamento** 
- Análise de Correlação Canônica

# Análise Multivariada de Dados

Unidades Amostras	Variáveis					
	1	2	...	j	...	p
1	$Y_{11}$	$Y_{12}$		$Y_{1j}$		$Y_{1p}$
2	$Y_{21}$	$Y_{22}$		$Y_{2j}$		$Y_{2p}$
...	...	...	...	...		...
i	$Y_{i1}$	$Y_{i2}$		$Y_{ij}$		$Y_{ip}$
...	...	...	...	...	...	...
n	$Y_{n1}$	$Y_{n2}$		$Y_{nj}$		$Y_{np}$

Análise no  $\mathbb{R}^{n \times n}$   
 Redução de dimensionalidade  
 das unidades amostrais!

Objetivos:

- **Formação de grupos de unidades amostrais  $\Rightarrow$  agrupamento de observações  $\Rightarrow$  grupos homogêneos internamente e heterogêneos externamente**
- A técnica também pode ser usada para identificar similaridades entre Variáveis  $\Rightarrow$  agrupamento de variáveis



**ANÁLISE DE AGRUPAMENTO (Cluster)**

# MOTIVAÇÃO

Cães pré-históricos da Tailândia (Manly, 2005).

Grupo	X1	X2	X3	X4	X5	X6
G1	9.7	21.0	19.4	7.7	32.0	36.5
G2	8.1	16.7	18.3	7	30.3	32.9
G3	13.5	27.3	26.8	10.6	41.9	48.1
G4	11.5	24.3	24.5	9.3	40.0	44.6
G5	10.7	23.5	21.4	8.5	28.8	37.6
G6	9.6	22.6	21.1	8.3	34.4	43.1
Cão Pré-h	10.3	22.1	19.1	8.1	32.2	35.0

Coord. Principais:

	$\hat{Y}_1$	$\hat{Y}_2$
1	-4.76	-0.28
2	-10.23	-2.78
3	13.90	0.18
4	8.26	-1.68
5	-3.98	4.29
6	2.01	0.00
7	-5.21	0.26

Permite  
visualizar a  
formação  
de grupos

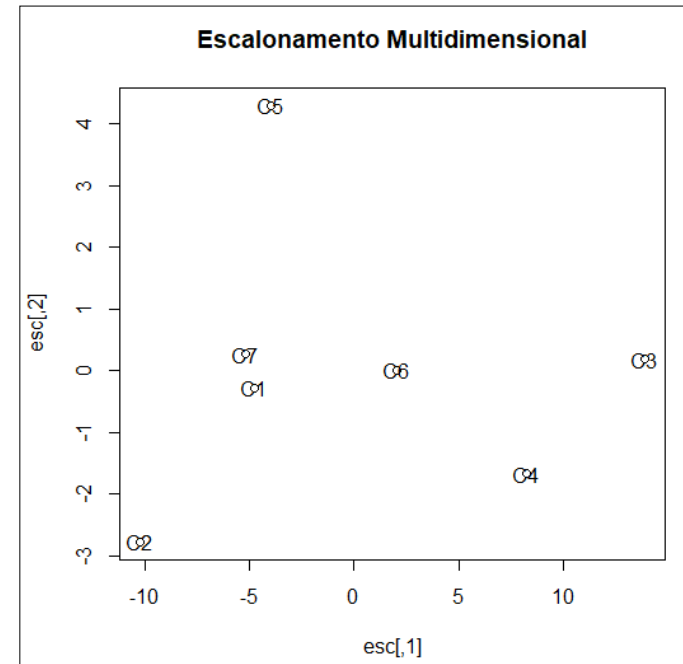
Matrizes de Distância

	1	2	3	4	5	6	7
1	0	6.21	18.70	13.13	4.83	7.43	2.03
2	6.01	0	24.34	18.55	9.44	12.94	6.62
3	18.67	24.31	0	5.99	18.38	12.50	19.20
4	13.10	18.52	5.94	0	13.64	7.26	13.78
5	4.64	9.44	18.35	13.62	0	7.98	5.09
6	6.78	12.55	11.89	6.48	7.36	0	8.67
7	0.70	5.87	19.11	13.61	4.21	7.22	0

D: Euclidiana

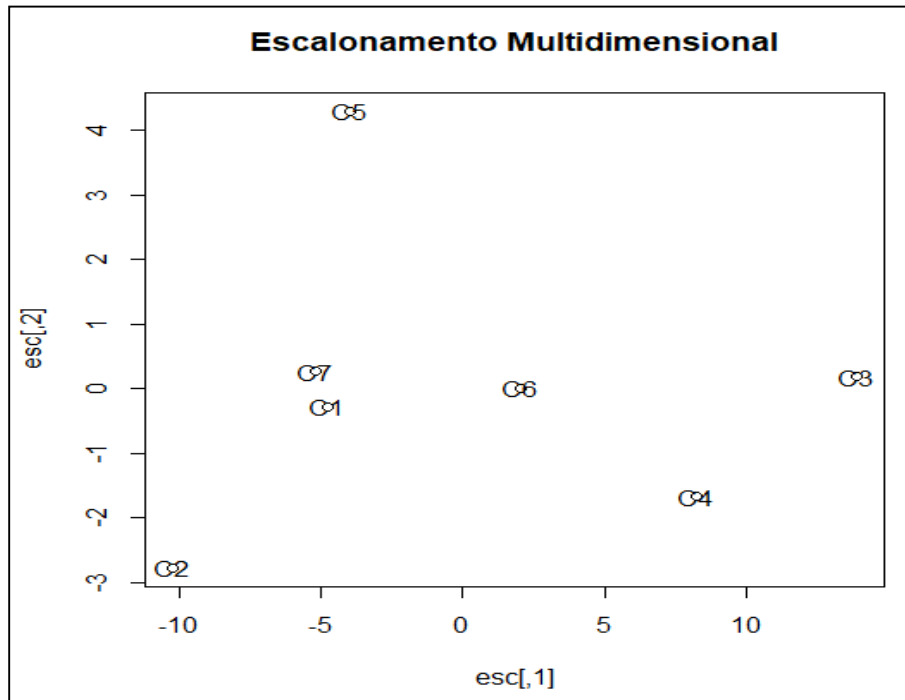
$= \hat{D}$

Representação das observações em  $\mathbb{R}^2$ : soluções equivalentes de **Componentes Principais** e **Escalonamento Multidimensional**

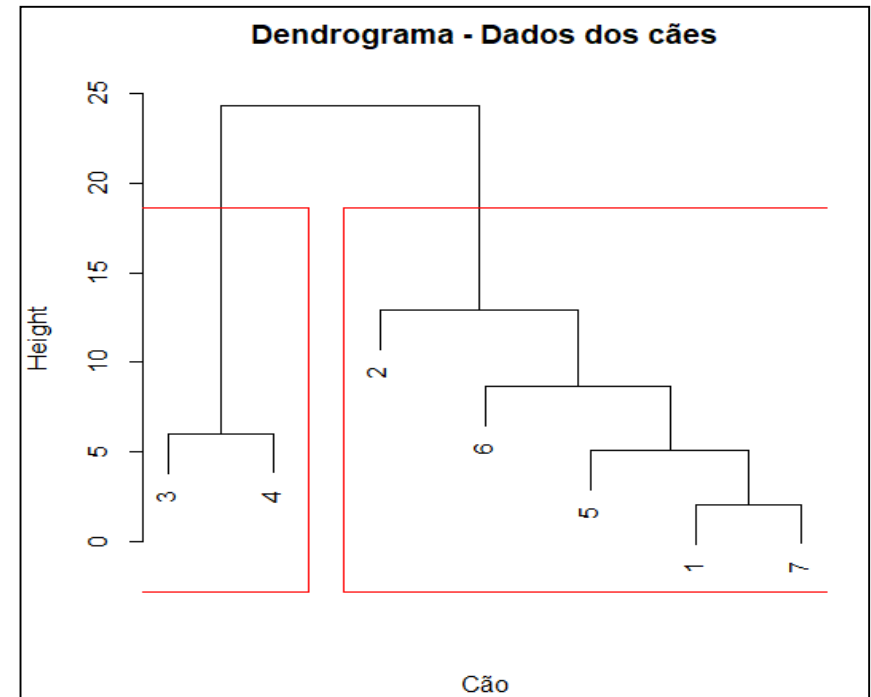


# Análise de Agrupamentos

## Escalonamento Multidimensional Coordenadas Principais



## Análise de Agrupamento (hierárquico)



Como estes agrupamentos foram formados?  
Ambas as análises foram realizadas à partir da matriz de distâncias  
(Euclidiana) entre as observações.

# Análise de Agrupamentos

Etapas da Aplicação de uma Análise de Agrupamento:

- **Escolha do Critério de Parecença**: adotar uma medida de distância (ou proximidade) entre pontos (uso de *variáveis originais ou padronizadas*).
- **Definição do Número de Grupos**: decisão a priori ou a posteriori (com base nos resultados da análise)
- **Formação dos grupos**: definir o algoritmo de formação dos grupos.
- **Validação do Agrupamento**: é comum supor que cada grupo seja uma amostra aleatória de uma subpopulação e aplicar técnicas inferenciais (comparações de médias dos grupos, por ex.). Algumas técnicas descritivas também são usadas (correlação cofenética e gráfico da silhueta).
- **Interpretação dos grupos**: caracterizar os grupos por meio de estatísticas descritivas e gráficos (radar, perfis de médias)

# Análise de Agrupamentos

## Medidas de Parecença

- Medidas de Dissimilaridade (Distância): quanto maior o valor, mais diferentes são os objetos.

### Variáveis quantitativas

$$d_{ik} = \sqrt{(Y_i - Y_k)'(Y_i - Y_k)} = \sqrt{\sum_{j=1}^p (Y_{ij} - Y_{kj})^2}$$

Distância Euclidiana entre observações  
(**originais ou padronizadas**)

$$d_{ik}^A = \sum_{j=1}^p |Y_{ij} - Y_{kj}|$$

Distância de Manhattan (quarteirão)

$$d_{ik}^M = \sqrt[m]{\sum_{j=1}^p |Y_{ij} - Y_{kj}|^m} ; m \geq 1$$

Distância de Minkowsky (mais geral)

⇒ Medidas de Parecença entre unidades amostrais também podem ser definidas para **dados Qualitativos** (Ex. distância de Bhattacharya para proporções) bem como para dados envolvendo variáveis Quantitativas e Qualitativas (ver, Johnson and Wichern)

# Formação de Grupos

Taxa de delitos (por 100.000 hab.) por divisão territorial de polícias do Estado de São Paulo (Deinter), em 2002\*.

Deinter	Homicídio doloso	Furto	Roubo	Roubo e furto de veículos
SJRP	10,85	1500,8	149,35	108,38
RP	14,13	1496,07	187,99	116,66
Bauru	8,62	1448,79	130,97	69,98
Campinas	23,04	1277,33	424,87	435,75
Sorocaba	16,04	1204,02	214,36	207,06
SP	43,74	1190,94	1139,52	909,21
SJC	25,39	1292,91	358,39	268,24
Santos	42,86	1590,66	721,9	275,89
Média	23,08	1375,19	415,92	298,9
dp	13,69	152,05	351,62	273,35

\*Artes, R. e Barroso, L.

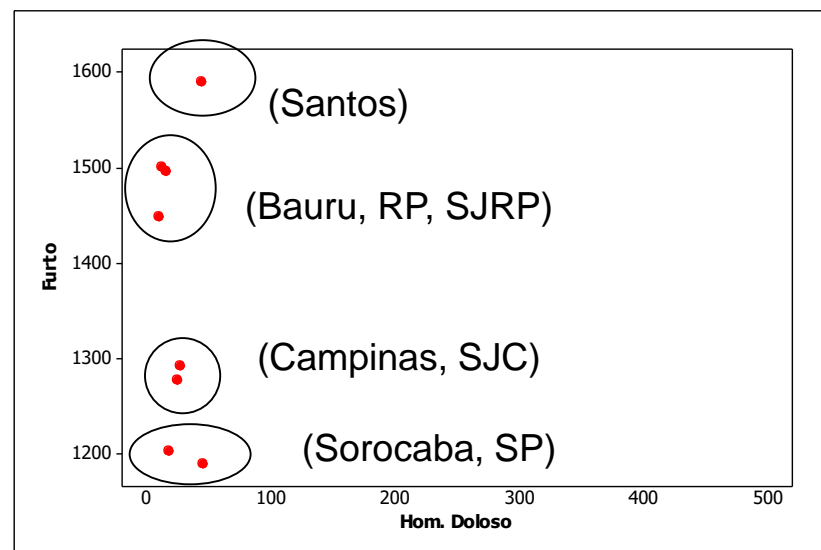
Como podemos agrupar as regiões?



# Formação de Grupos

Como podemos agrupar as regiões que sejam homogêneas quanto à incidência de homicídio doloso e furto?

Deinter	Homicídio doloso	Furto	Roubo	Roubo e furto de veículos
SJRP	10,85	1500,8	149,35	108,38
RP	14,13	1496,07	187,99	116,66
Bauru	8,62	1448,79	130,97	69,98
Campinas	23,04	1277,33	424,87	435,75
Sorocaba	16,04	1204,02	214,36	207,06
SP	43,74	1190,94	1139,52	909,21
SJC	25,39	1292,91	358,39	268,24
Santos	42,86	1590,66	721,9	275,89
Média	23,08	1375,19	415,92	298,9
dp	13,69	152,05	351,62	273,35



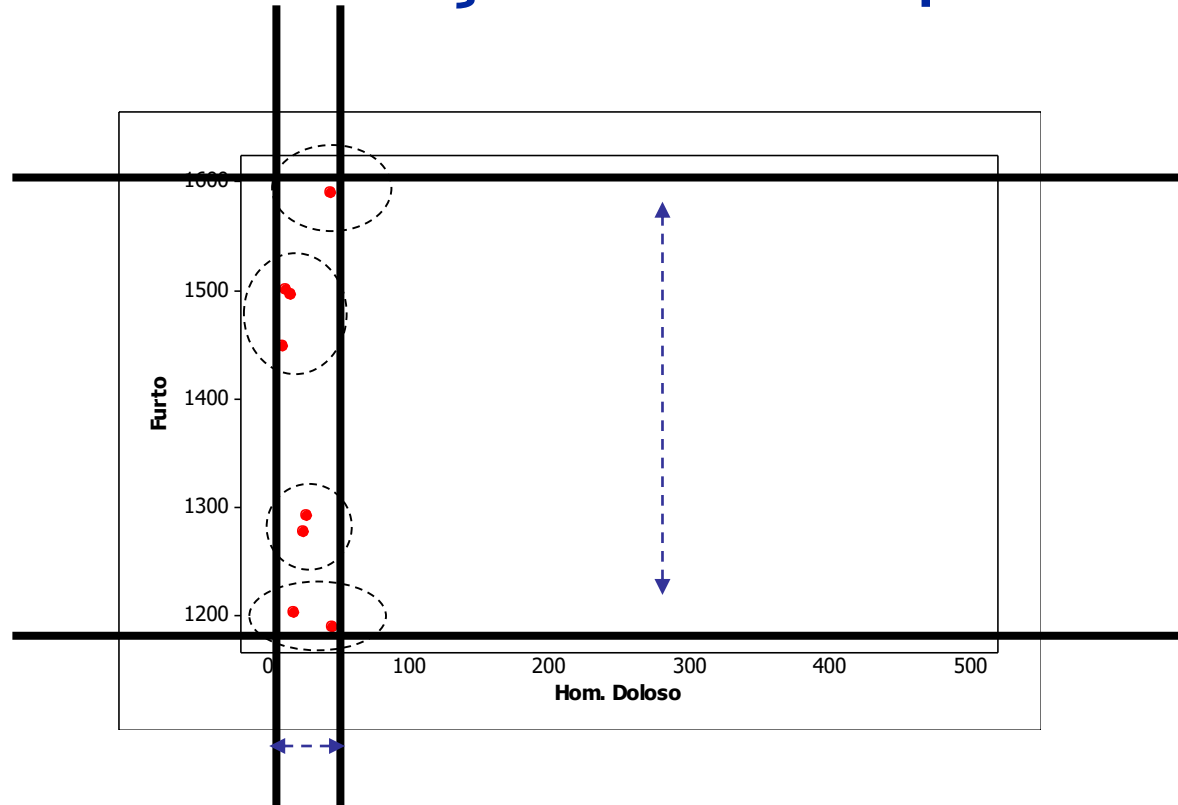
Mesma escala!

Formação de 4 grupos

Considerando duas variáveis  $\Rightarrow$  Uso de Diagrama de Dispersão

Alternativa 1: Critério de Formação dos Grupos  $\Rightarrow$  considerar a “proximidade” entre os pontos na escala original.

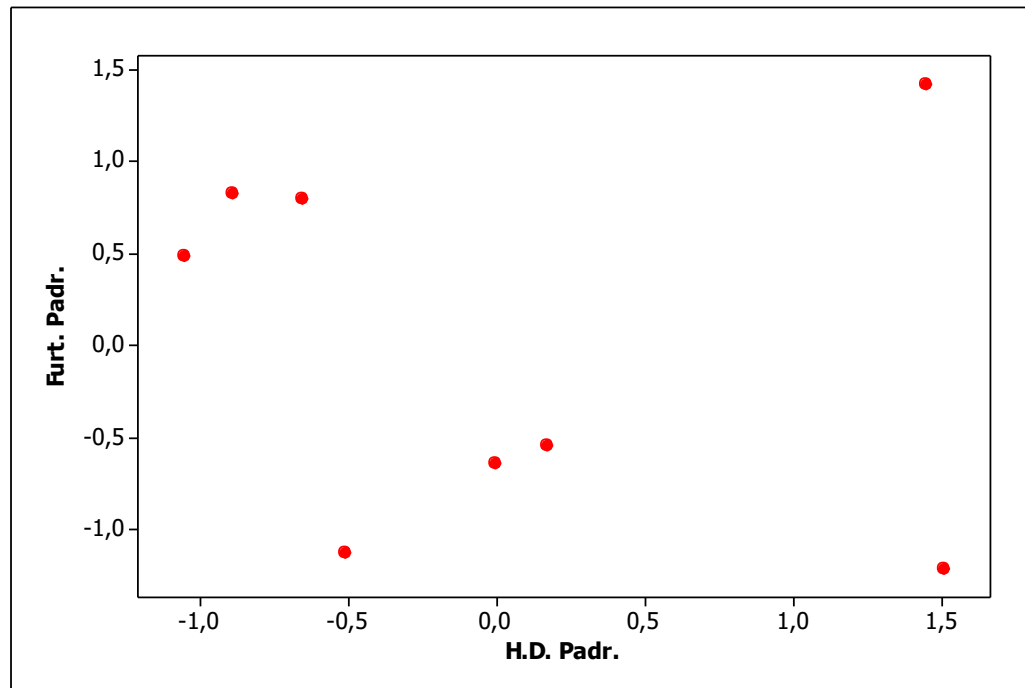
# Formação de Grupos



- Critério de Proximidade dos pontos (por inspeção visual)  $\Rightarrow$  as distâncias no sentido vertical são muito maiores do que no sentido horizontal  $\Rightarrow$  devido à variabilidade da variável furto ser maior que a de homicídio doloso
- A var. Homicídio doloso contribuiu pouco para a formação dos grupos  $\Rightarrow$  adotar critérios que atribuam igual importância às variáveis (**padronização**)

# Formação de Grupos

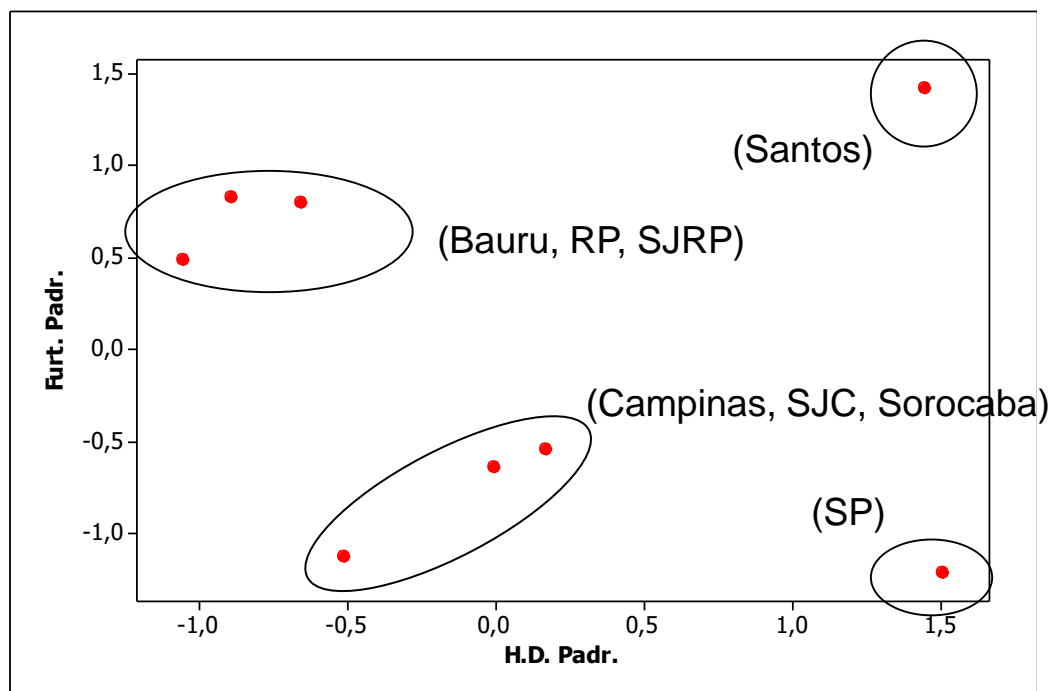
Uso de variáveis PADRONIZADAS  $\Rightarrow$  diagrama de dispersão mostra que as distâncias no sentido vertical e horizontal são da mesma grandeza  $\Rightarrow$  as duas variáveis estão recebendo importância equivalente.



Formar quatro grupos homogêneos de regiões!

# Formação de Grupos

Uso de variáveis PADRONIZADAS  $\Rightarrow$  diagrama de dispersão mostra que as distâncias no sentido vertical e horizontal são de mesma grandeza  $\Rightarrow$  as duas variáveis estão recebendo importância equivalentes



# Análise de Agrupamentos

## Outras Medidas de Parecença

- Variáveis Quantitativas: pode-se utilizar o **coeficiente de correlação de Pearson** como medida de parecença entre pares de unidades amostrais  $\Rightarrow$  quanto mais próximo de 1 (ou -1) maior a similaridade e quanto mais próximo de 0 maior a dissimilaridade.

$\Rightarrow$  Transformar a correlação em uma medida de dissimilaridade

$$d_{ii'} = (r_{ii} + r_{i'i'} - 2r_{ii'})^{1/2}$$

- Nem sempre faz sentido adotar a correlação como medida de parecença “entre unidades amostrais” (Ex. Unidades amostrais avaliadas em variáveis de diferentes unidades de medidas)
- O coeficiente de correlação  $r$  é comumente usado como medida de “parecença” entre variáveis (e não entre unidades amostrais)
- A correlação entre unidades amostrais valoriza padrões de forma (tendências, como em dados longitudinais) e a distância valoriza mais padrões de tamanho

# Análise de Agrupamentos

## Algoritmos de Agrupamento

- **Métodos Hierárquicos Aglomerativos**: os agrupamentos hierárquicos partem dos objetos individuais ( $n$ ) para a formação de um único grupo.
  - Método do Vizinho mais Próximo/Perto (Ligação Simples)
  - Método do Vizinho mais Distante/Longe (Ligação Completa)
  - Método das Médias das Distâncias (Ligação Média)
  - Método da Centróide
  - Método de Ward
- **Métodos de Partição**: os agrupamentos não hierárquicos buscam a partição de  $n$  objetos em  $K$  grupos.
  - Algoritmo das K-Médias

# Análise de Agrupamentos

## Algoritmos de Agrupamentos Hierárquicos

- **Método do Vizinho mais Distante (Ligação Completa ou Distância Máxima)**: a distância entre os grupos  $G_1$  e  $G_2$  é dada pela maior distância entre os elementos de cada grupo

$$d(G_1, G_2) = \max_{i \in G_1, k \in G_2} d_{ik} \quad \Rightarrow \text{Forma grupos de alta homogeneidade interna}$$

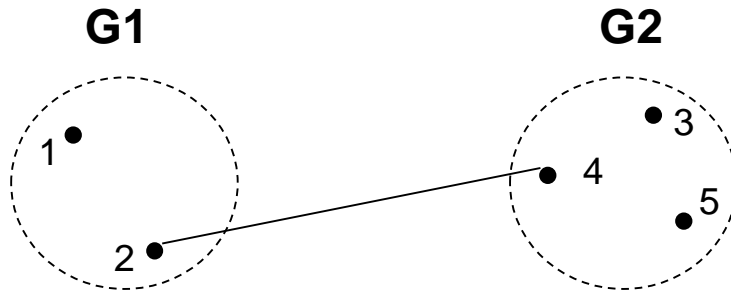
- **Método do Vizinho mais Perto (Ligação Simples ou Distância Mínima)**: a distância entre os grupos  $G_1$  e  $G_2$  é dada pela menor distância entre os elementos de cada grupo

$$d(G_1, G_2) = \min_{i \in G_1, k \in G_2} d_{ik} \quad \Rightarrow \text{Pode não distinguir grupos pobremente separados}$$

- **Método das Médias das Distâncias (Ligação Média)**: a distância entre os grupos é obtida pelo cálculo da média das distâncias entre os elementos de cada grupo

$$d(G_1, G_2) = \frac{\sum_i \sum_k d_{ik}}{n_i n_k}$$

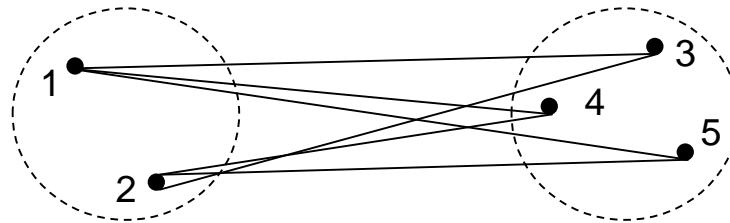
# Algoritmos Hierárquicos



Distância entre Grupos

**Ligação Simples**

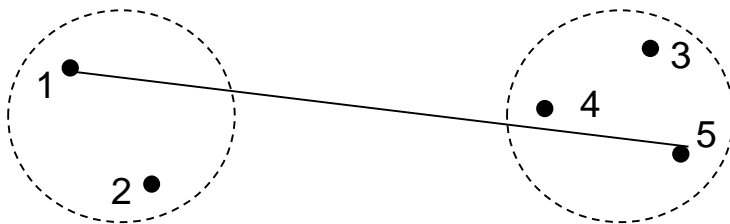
$$d(G_1, G_2) = d_{24}$$



**Ligação Média**

$$d(G_1, G_2) = \frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{6}$$

6 ← 2x3



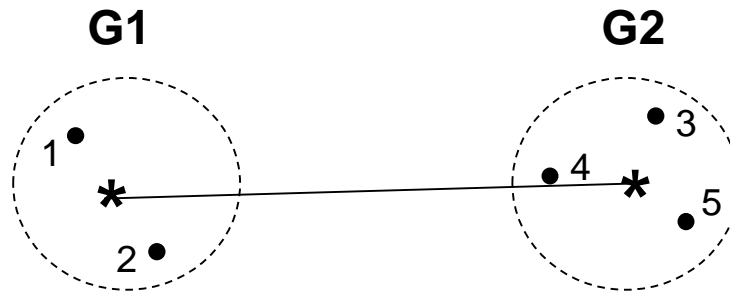
**Ligação Completa**

$$d(G_1, G_2) = d_{15}$$



# Algoritmos Hierárquicos

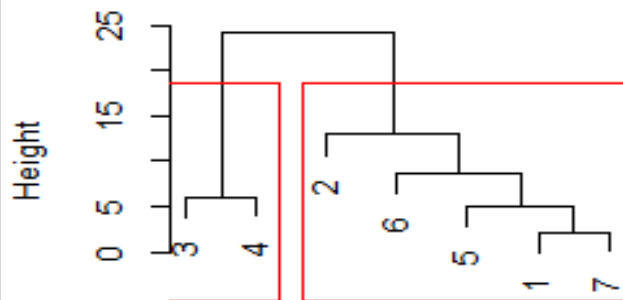
- **Método de Centróides**; este método define a coordenada de cada grupo como sendo a média das coordenadas de seus elementos. Uma vez obtida esta coordenada comum (denominada centróide) a distância entre os grupos  $G_1$  e  $G_2$  é dada pela distância entre as centróides.



# Análise de Agrupamento – Dados dos Cães

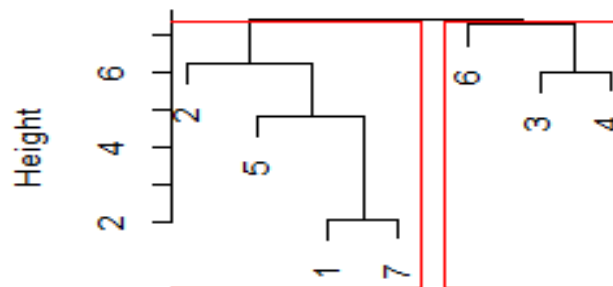
p=6

**Dendrograma - Dados dos cães**



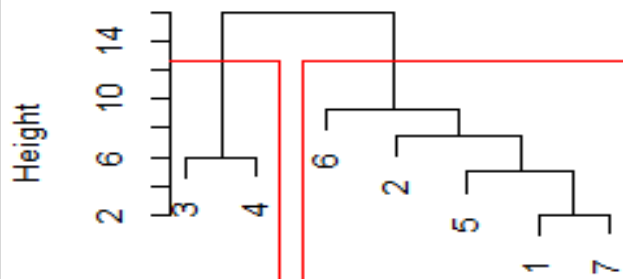
Cão  
Lig. Completa - Dist. Euclidiana

**Dendrograma - Dados dos cães**



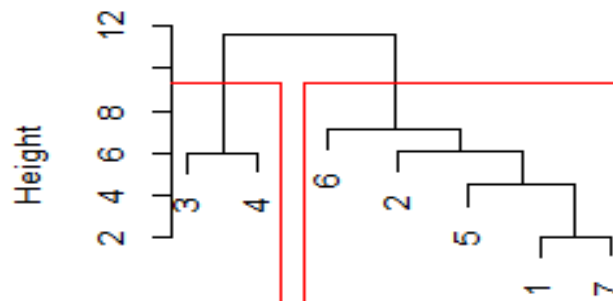
Cão  
Lig. Simples - Dist. Euclidiana

**Dendrograma - Dados dos cães**



Cão  
Lig. Média - Dist. Euclidiana

**Dendrograma - Dados dos cães**



Cão  
Lig. Centróide - Dist. Euclidiana

Adotar um algoritmo de agrupamento

Estabelecer um ponto de corte no dendrograma (eixo y) para a formação de grupos!

Neste caso, formar 2 grupos de cães.

# Algoritmos Hierárquicos

- **Método de Ward**: é atraente pelo forte apelo estatístico envolvido. Busca formar grupos com máxima homogeneidade interna (DENTRO) e máxima heterogeneidade externa (ENTRE). O procedimento baseia-se na decomposição da Soma de Quadrados Total de uma Análise de Variância (ANOVA).

Considere a formação de **L grupos** de observações, especificamente, por meio de valores da **variável Y1**.

Y1			
G1	G2	...	GL
-	-		-
-	-	$Y_{i1}$	-
-			-
-			-
$\bar{Y}_{G_1 1}$	$\bar{Y}_{G_2 1}$		$\bar{Y}_{G_L 1}$

Como particionar a soma de quadrados total em componentes de variabilidade ENTRE e DENTRO de grupos?

$\bar{Y}_1$  Lembre da ANOVA!

# Algoritmos Hierárquicos

Y1			
G1	G2	...	GL
-	-		-
-	-	$Y_{il}$	-
-	-		-
-	-		-
$\bar{Y}_{G_1}$	$\bar{Y}_{G_2}$		$\bar{Y}_{G_L}$
$n_{G_1}$	$n_{G_2}$		$n_{G_L}$
$\bar{Y}_1$			

Variável Y1

p = 1

$$SQT(1) = SQE(1) + SQD(1)$$

**ANOVA**

$$\sum_{l=1}^L \sum_{i \in G_l} (Y_{il} - \bar{Y}_1)^2 = \sum_{l=1}^L n_{G_l} (\bar{Y}_{G_l} - \bar{Y}_1)^2 + \sum_{l=1}^L \sum_{i \in G_l} (Y_{il} - \bar{Y}_{G_l})^2$$



Método de Ward  $\Rightarrow$  Minimizar SQD (soma de quadrados dentro) e maximizar SQE (soma de quadrados entre)

# Algoritmos Hierárquicos

**Método de Ward:** Para considerar as **p variáveis** simultaneamente a Soma de Quadrados (Dentro) da Partição é definida como:

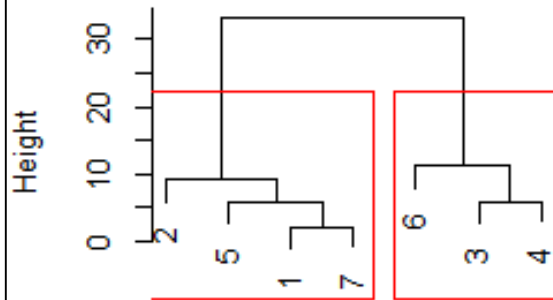
$$SQDP = \sum_{j=1}^p SQD(j)$$

## Procedimento:

- **Passo 1:** Para **n pontos**, calcular SQDP para os possíveis **(n-1)** grupos distintos e selecionar o agrupamento com a menor SQDP ( $\exists C_2^n$ )
- **Passo 2:** Calcular SQDP para os possíveis **(n-2)** grupos distintos (fixada a união obtida no Passo 1) e selecionar o agrupamento com a menor SQDP
- Os próximos passos consistem na formação de **(n-3), (n-4), ..., 1** grupos, selecionando-se sempre o agrupamento com menor SQDP
- O número de grupos é definido em função dos saltos em cada passo.

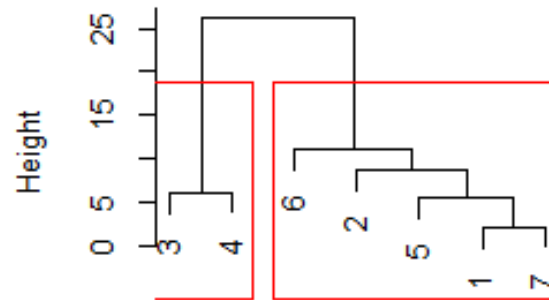
# Análise de Agrupamentos – Dados dos Cães

Dendrograma - Dados dos cães



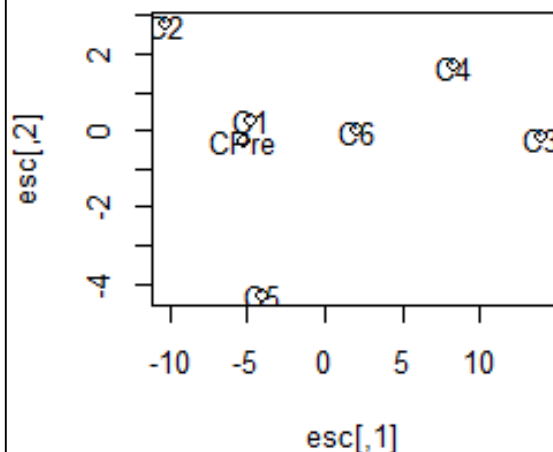
Cão  
Ward.D - Dist. Euclidiana

Dendrograma - Dados dos cães



Cão  
Ward.D2 - Dist. Euclidiana

Escalonamento Multidimensional



O **cão C6** deve pertencer  
a qual agrupamento?

No R, a função  
“hclust” oferece dois  
algoritmos diferentes  
que implementam o  
método de Ward:  
“ward.D” e  
“ward.D2”

Escolher o que  
oferecer melhor  
interpretação.  
O gráfico das  
Coordenadas  
Principais pode  
ajudar na decisão  
pois permite a  
avaliação visual no  
gráfico  
bidimensional!

# Método das K-Médias

## K-Means: Método de Partição (Não-Hierárquico)

- Passo 1: Formação de uma partição inicial. Em geral, adota-se k observações como sementes do algoritmo para formação de k grupos.
- Passo 2: Percorrer a lista de observações e calcular as distâncias de cada uma delas ao CENTRÓIDE (médias) do grupo. Fazer a re-alocação da observação ao grupo em que ela apresentar menor distância. Re-calcular os centróides dos grupos que ganharam e perderam observações.
- Passo 3: Repetir o Passo 2 até que nenhuma alteração seja feita.
- Passo 4: Adotar uma função objetivo e, em cada passo, calcular seu valor para avaliação da partição. Identificar novas mudanças na formação dos grupos que possam otimizar ainda mais a função objetivo.

Funções objetivo mais comuns a serem minimizadas:

SQDP (Soma de Quadrados Dentro da Partição)

Distância Euclidiana ao quadrado das observações ao centróide

# Método das K-Médias

Implementado no R

## Algoritmo de Lloyd (ou Forgy):

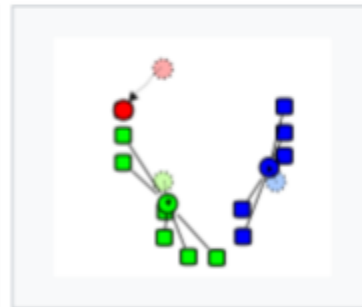
- Estabelecer  $K$  observações como **centróides** iniciais dos grupos, de forma aleatória.
- Atribuir cada uma das observações ao grupo cuja sua **distância** em relação ao centróide é a menor, entre todos os  $K$  centróides calculados.
- Quando todas as observações forem alocadas a algum grupo, **recalcular** os  $K$  centróides.
- Repetir os dois passos anteriores até que os centróides não sofram mais alterações (ou até um número máximo de iterações).



1.  $k$  initial "means" (in this case  $k=3$ ) are randomly generated within the data domain (shown in color).



2.  $k$  clusters are created by associating every observation with the nearest mean. The partitions here represent the **Voronoi diagram** generated by the means.



3. The **centroid** of each of the  $k$  clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.



# Método das K-Médias

Implementado no R  
(default)

Algoritmo de Hartigan Wong (1979):

- Fazer uma **partição aleatória inicial** das  $n$  observações em  $K$  grupos.
- Selecionar uma observação, de forma aleatória, **removê-la** do seu grupo e recalculando o respectivo centróide.
- Realocar a observação removida em algum dos grupos, de forma a **minimizar** a quantidade  $D$ . Recalculando o respectivo centróide.
- Repetir os dois passos anteriores até a convergência da  $D$ , que é necessariamente decrescente nesse processo.



Procedimento K-Médias++ (Arthur e Vassilvitskii, 2007): seleção alternativa das sementes na partição inicial, de forma a garantir maior “espalhamento” dos grupos formados

# Método das Partições: K-Médias

## Dados dos Cães

**Grupos (K=2) - Algorithm "Hartigan-Wong"**

C1	C2	C3	C4	C5	C6	C7
2	2	1	1	2	2	2

**Tamanho dos Grupos:** 2 5

**Centróides dos grupos por variável**

	X1	X2	X3	X4	X5	X6
1	12.50	25.80	25.65	9.95	40.95	46.35
2	9.68	21.18	19.86	7.92	31.54	37.02

**Soma de Quadrados QTotal:** 481.72

**Soma de Quadrados Dentro de Grupos:** 17.920 117.316

**Soma de Quadrados Entre Grupos:** 346.484      **SQE/SQTotal = 71.9%**

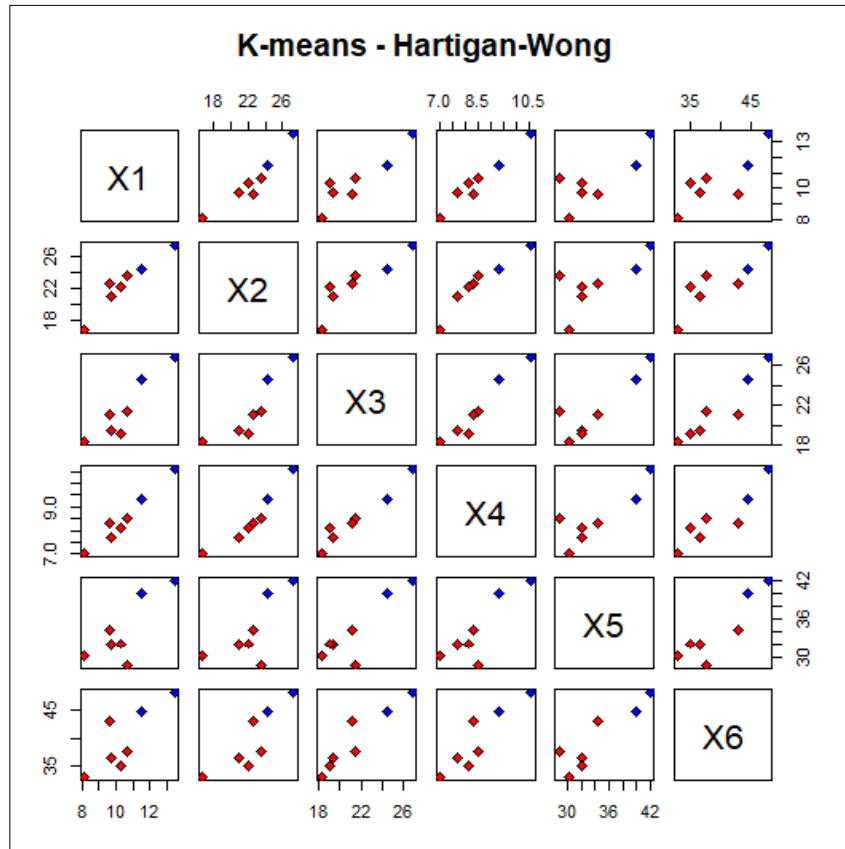
**Algoritmo "Lloyd":**

C1	C2	C3	C4	C5	C6	C7
1	1	2	2	1	2	1

1 1 2 2 1 2 1      SQE/SQTotal = 71.4%

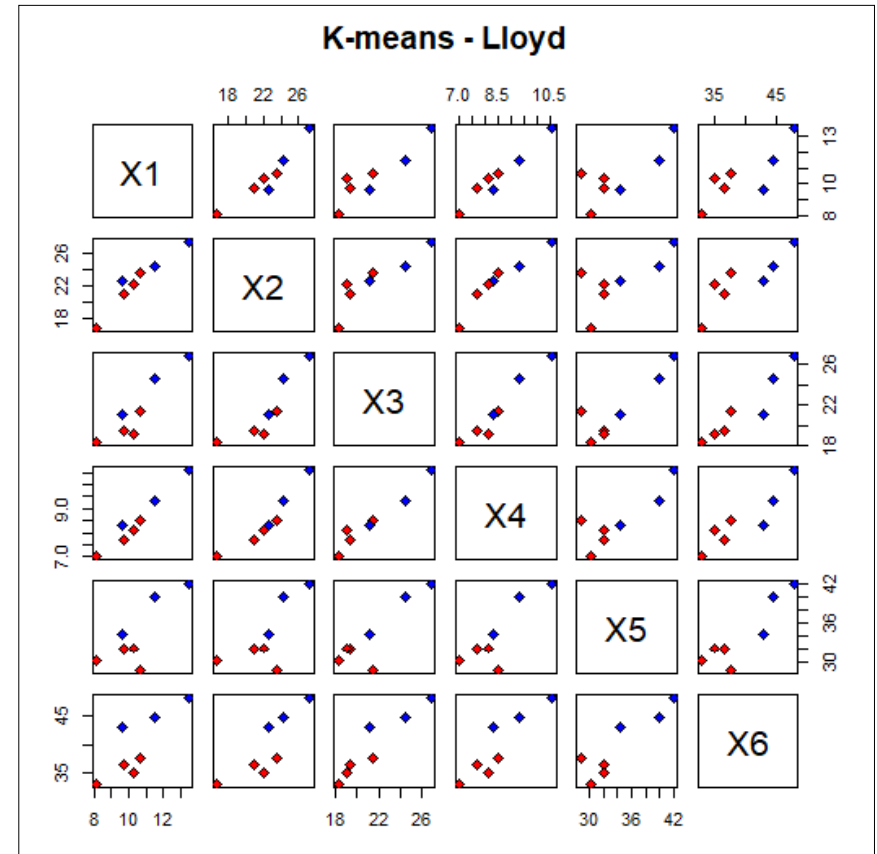
# Método das Partições: K-Médias

## Avaliação dos Grupos na Matriz de Gráficos de Dispersão



Agrupamentos

C1	C2	C3	C4	C5	C6	C7
1	1	2	2	1	1	1



Agrupamentos

C1	C2	C3	C4	C5	C6	C7
1	1	2	2	1	2	1

# Análises Supervisionadas e Não-Supervisionadas

Medidas biométricas (mm) de Pardais fêmea

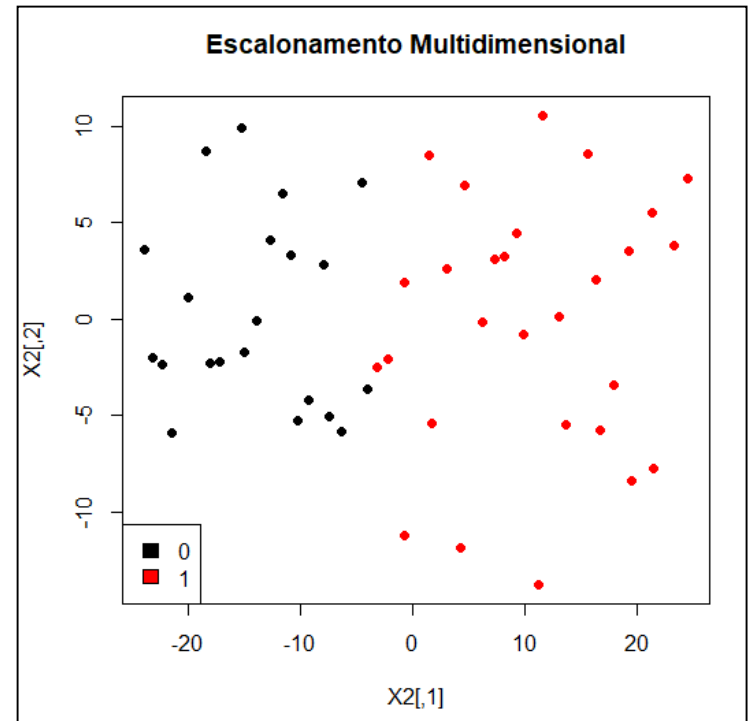
(Manly, 2005; Hermon Bumps, 1898).

Pardal	Sobrev.	X1	X2	X3	X4	X5
1	S	156	245	31.6	18.5	20.5
...	...					
21	S	159	236	31.5	18.0	21.5
22	N	155	240	31.4	18.0	20.7
...	...					
49	N	164	248	32.3	18.8	20.9

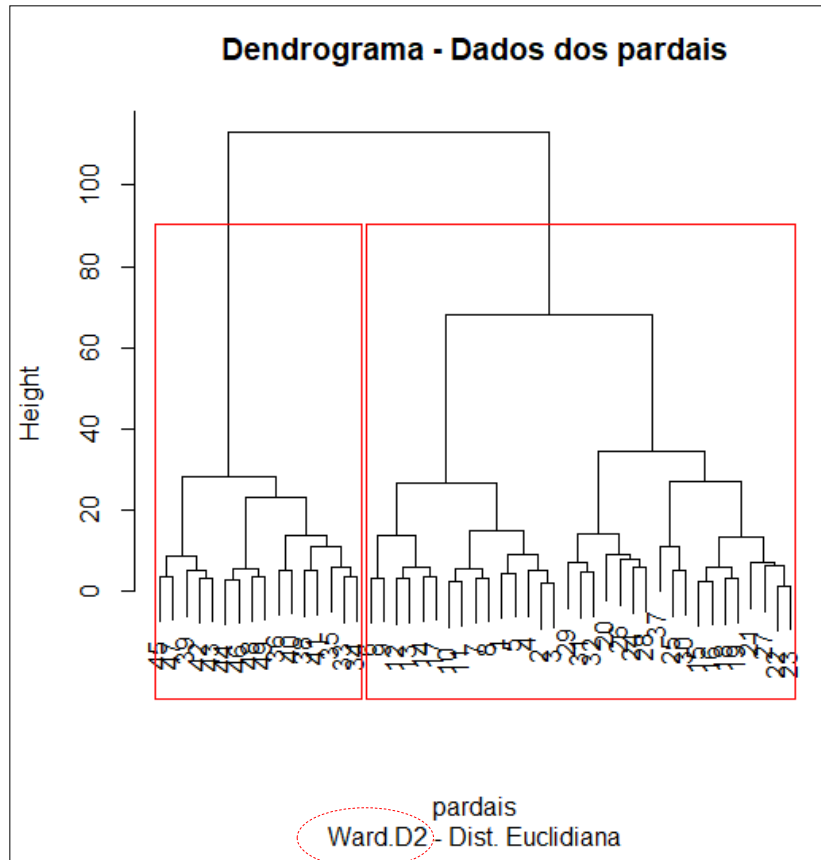
Como os grupos de pardais Sobreviventes (=0) e Não Sobreviventes (=1) podem ser preditos?  $\Rightarrow$  **Análise Discriminante (técnica supervisionada)!**

As variáveis biométricas (X1 a X5) permitem a classificação dos pardais em Sobreviventes (=0) e Não Sobreviventes (=1)?  $\Rightarrow$  **Análise de Escalonamento**

**Multidimensional e Análise de Agrupamento (k=2) são técnicas não-supervisionadas (não levam em conta os grupos) construídas a partir de uma matriz de distâncias**



# Análise de Agrupamento



```
Pred
grup 1 2
      0 21 0
      1 12 16
%correta: 75.5
```

## Algoritmo k-Means Lloyd

```
Pred
grup 1 2
      0 13 8
      1 12 16
%correta: 59.2
```

## Algoritmo k-Means Hartigan-Wong

```
Pred
grup 1 2
      0 13 8
      1 12 16
%correta: 59.2
```

## Algoritmo kNN(3) K-Vizinhos mais Próximos

```
Pred
grup 1 2
      0 22 1
      1 6 20
%correta: 85.7
```

Considere  
também:



# Método de k Vizinhos mais Próximos – kNN

## k-nearest neighbors algorithm

- kNN é uma **Técnica de Predição** na qual, para qualquer ponto Y que desejamos prever, é construída uma “vizinhança” com os k pontos mais próximos de Y, e então uma “**média**” (ou grupo mais votado) desses pontos é tomada como estimativa.
- **kNN é um dos algoritmos**, dentre muitos, utilizado para prever Y.
- De maneira geral, problemas de Predição podem ser definidos com Y uma variável contínua ou categórica, sendo que, independentemente do tipo da variável Y, a função de predição é baseada na **Esperança Condicional de Y dado Preditores X**:

Y categórico (binário):  $f(x) = P(Y = 1 | X = x) = E(Y | X = x)$

Y contínuo:  $f_{Y|X}(x) = E(Y | X = x)$

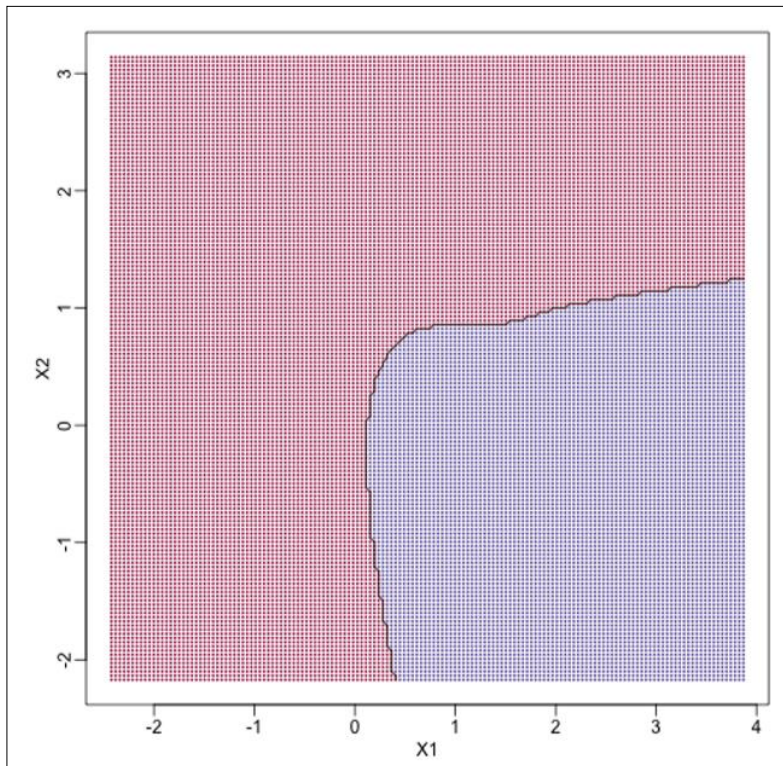
- Em geral, a função de predição é selecionada por minimizar a distância esperada entre o preditor e Y:

$$res = E\left(\left(\hat{Y} - Y\right)^2 | X = x\right)$$

# Método de k Vizinhos mais Próximos - kNN

- Problemas de Predição têm sido abordados no escopo de “*Machine Learning*”, em que, algoritmos baseados no ajuste de modelos de regressão, bem como no método kNN, são bastante utilizados para finalidade de predição.

$$f(x_1, x_2) = E(Y | X_1 = x_1, X_2 = x_2)$$



Exemplo de função de predição simulada (Irizarry e Love, 2015), tal que:

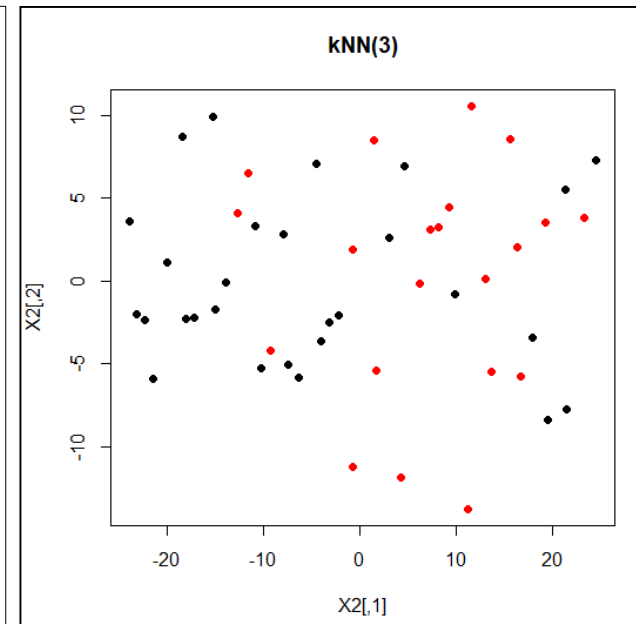
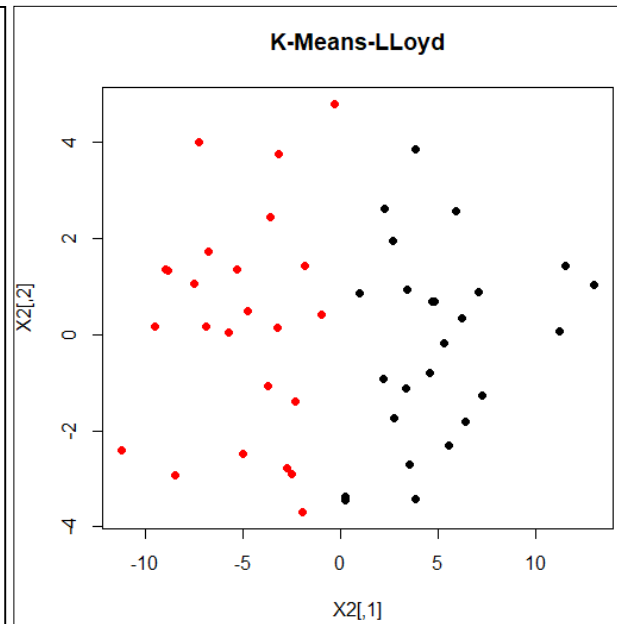
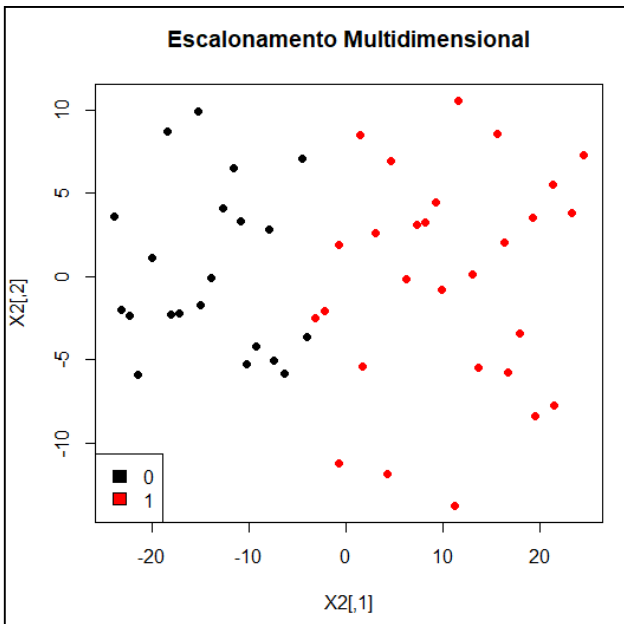
$$E(Y | x_1, x_2) \geq 0.5 \rightarrow \text{Grupo 1 (vermelho)}$$

$$E(Y | x_1, x_2) < 0.5 \rightarrow \text{Grupo 2 (azul)}$$

O **algoritmo kNN** é assim uma técnica de Agrupamento de observações bastante flexível. Neste caso, o **kNN** encontra a função de predição em um problema de predição por meio do ajuste de curvas suavizadas (do tipo Loess).

# Método de k Vizinhos mais Próximos - kNN

Dados dos Pardais:



Escalonamento  
Multidimensional  
(não-supervisionado)

**Algoritmo k-Means  
Lloyd**

```
      Pred
grup  1  2
    0 13  8
    1 12 16
%correta: 59.2
```

**kNN com k=3**

**Algoritmo kNN(3)  
K-Vizinhos mais  
Próximos**

```
      Pred
grup  1  2
    0 22  1
    1  6 20
%correta: 85.7
```



# Método de k Vizinhos mais Próximos - kNN

Dados dos Pardais

$n=21+28=49$

Amostras de Treinamento e  
Teste iguais

**kNN (k=1)**

```
Pred
true  0  1
      0 21  0
      1  0 28
%correta: 1.0
```

**kNN (k=3)**

```
Pred
grup  0  1
      0 22  1
      1  6 20
%correta: 85.7
```

**kNN (k=49)**

```
Pred
true  0  1
      0  0 21
      1  0 28
%correta: 57.1
```

**kNN usado junto com Validação Cruzada**

$n=21+28=49$

**kNN (k=3)**

**CV com K=10 Folds** (subamostras balanceadas)

Fold	01	02	03	04	05	06	07	08	09	10
	5	5	5	5	5	4	5	5	5	5

←  $n=49$

- Amostra de Treinamento: **Dados (-Fold01)**
- Aplicar **kNN (k=3)**
- Amostra Teste: **Dados (Fold01)**

	Pred		
true	0	1	
0	1	2	
1	1	2	

	Fold	
[1]	"1)	error rate: 0.5"
[1]	"2)	error rate: 0"
[1]	"3)	error rate: 0.4"
[1]	"4)	error rate: 0.6"
[1]	"5)	error rate: 0.167"
[1]	"6)	error rate: 0.2"
[1]	"7)	error rate: 0.25"
[1]	"8)	error rate: 0.5"
[1]	"9)	error rate: 0"
[1]	"10)	error rate: 0.5"

# kNN combinado com Validação Cruzada

kNN (k=1)    kNN (k=2)    ...    kNN (k=49)

↓ k=1

## Validação Cruzada

n=21+28=49

K=10 Folds (balanceados)

Fold	01	02	03	04	05	06	07	08	09	10
	5	5	5	5	5	4	5	5	5	5

K = 1, ..., 10

↓ K=1

Amostra Treinamento: Dados(-Fold(k))

Amostra Teste: Dados(Fold(k))

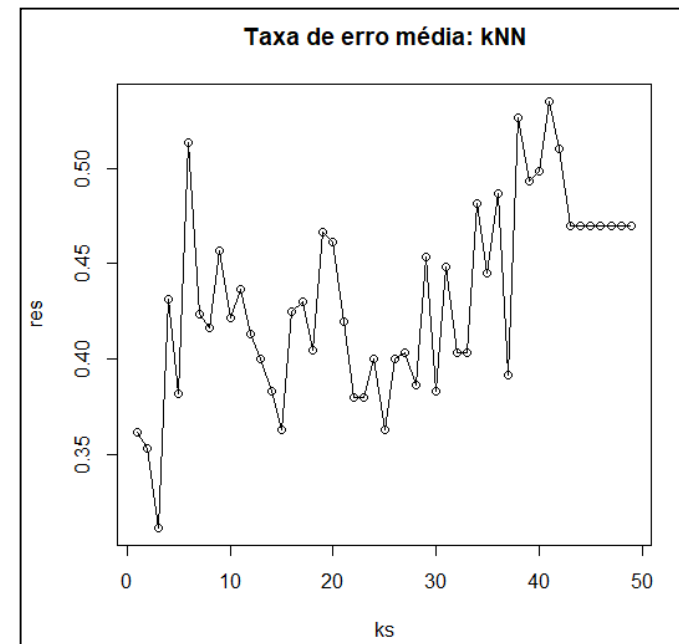
%Classificação errada

Média da %Classificação errada

↓ K=K+1

↓ k=k+1

Dados dos Pardais



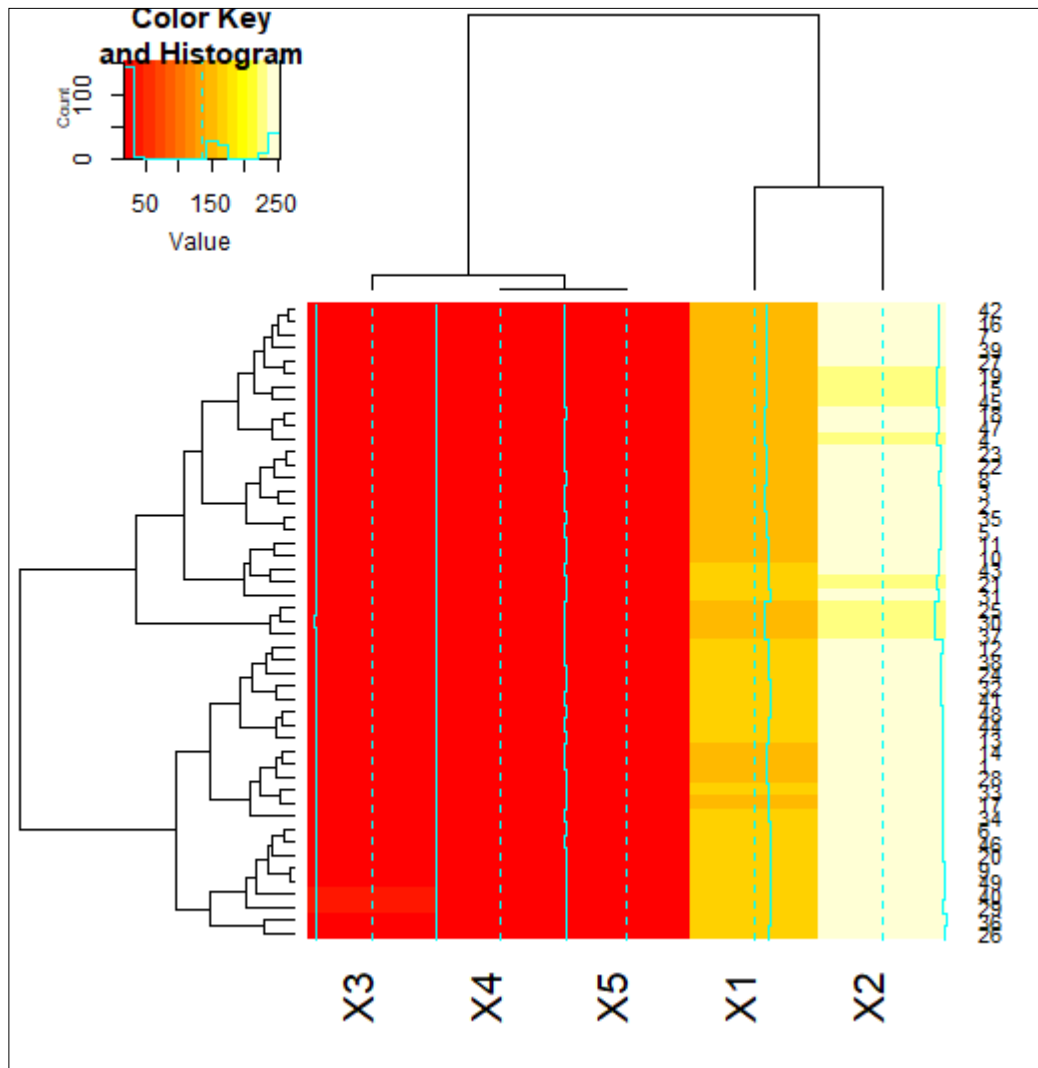
kNN com k=1, 2,...,49

CV com K=10

%Média Erro classificação=0.312

# Análise de Agrupamento

## Aplicação: Heatmap



### Dados dos Pardais

**Heatmap:** Representação simultânea dos agrupamentos das unidades amostrais e das variáveis

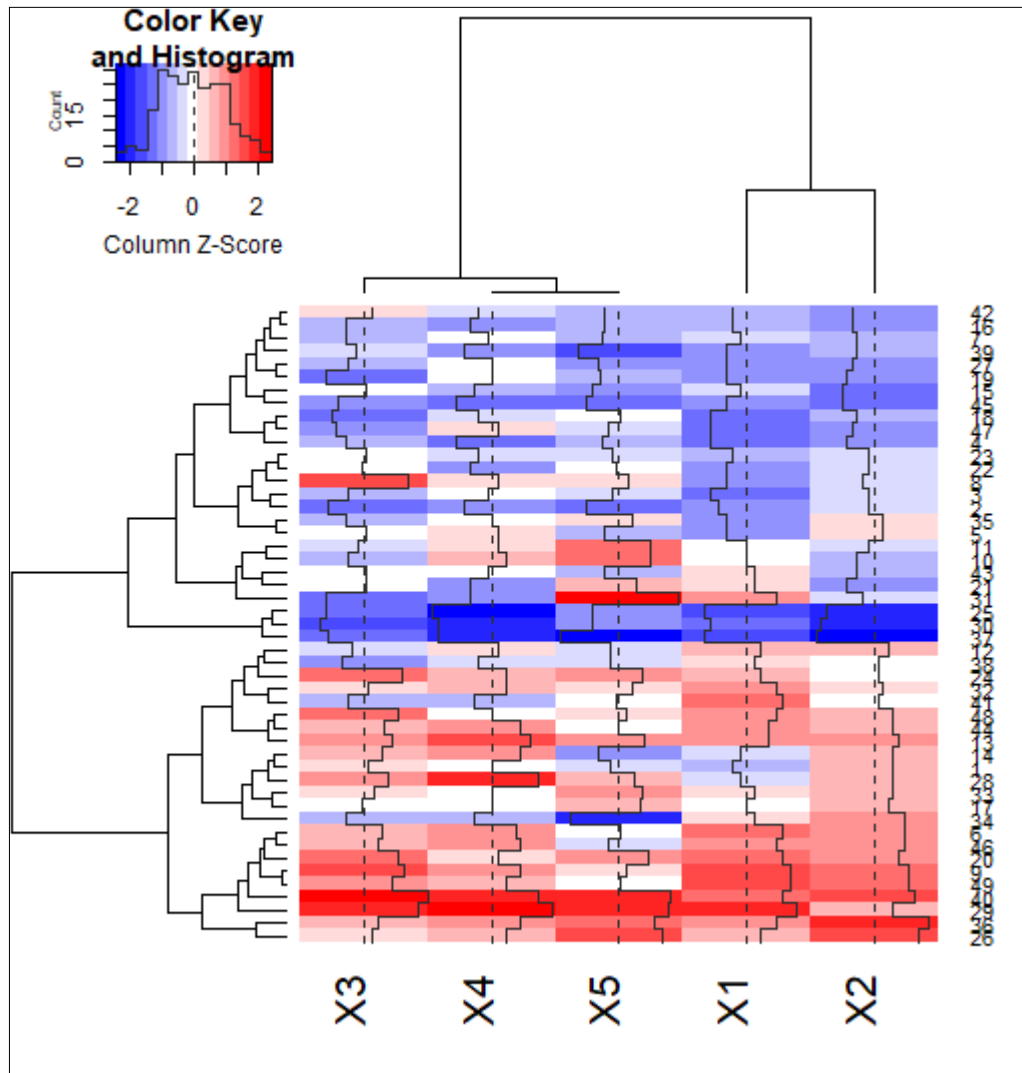
Agrupamento hierárquico (Ligação Completa) das unidades amostrais e das variáveis.

O valor das variáveis define as cores no gráfico.

As variáveis X1 e X2 são as que mais discriminam os grupos formados, os quais são homogêneos para as demais variáveis X3, X4 e X5.

# Análise de Agrupamento

## Aplicação: Heatmap



### Dados dos Pardais

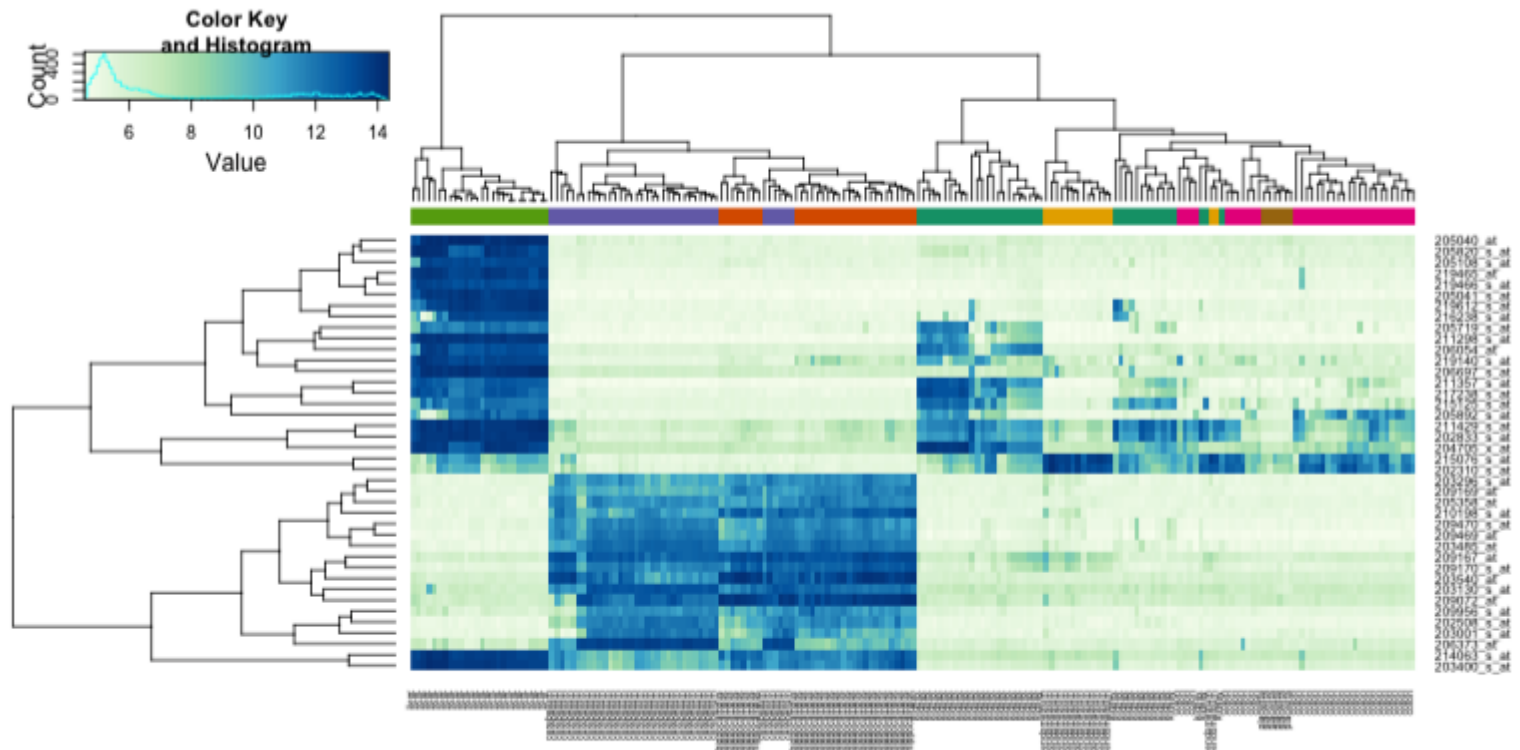
Agrupamento hierárquico (Ligação Completa) das unidades amostrais e das variáveis.

Os escores z das colunas definem as cores.

Para todas as variáveis, supondo a formação de dois grupos de unidades amostrais, o primeiro é caracterizado pelos maiores valores do escore z para todas as variáveis.

# Análise de Agrupamento

## Aplicação: Heatmap



Heatmap created using the 40 most variable genes and the function heatmap.2.

Irizarry and Love (2015)

## Dados de expressão gênica log-transformados (cores)

Linhas: representação de 40 genes

Colunas: representação de 189 amostras (tecidos cancerosos)