Introdução a Métodos Estatísticos para a Bioinformática

Profa. Júlia Maria Pavan Soler

pavan@ime.usp.br

IBI 5086 – Bioinformática - IME/USP

2º Sem/2023

Programa

- ➢ Álgebra linear básica: cálculo matricial, determinantes, sistemas lineares, produto interno, norma, ortogonalidade, autovalores e autovetores
- ✓ Estrutura de Dados: variáveis (resposta, explicativa), unidades amostrais e experimentais

- ✓ 1.1. Comparação de 2 ou mais grupos: Testes Clássicos (teste t, Wilcoxon, ANOVA), Testes de Aleatorização, Comparações Múltiplas, Efeitos Genéticos
- ✓ 1.2. Análise de Tabelas de Contingência: Testes Qui-Quadrado, Regressão Logística.

2. **Análise Multivariada de Dados**: Componentes Principais, Coordenadas Principais, Análise de Correspondência, Análise Discriminante (MANOVA), **Análise de Agrupamento,** Correlação Canônica

3. Simulação de Monte Carlo, Intervalos de Confiança Bootstrap

Análise de Agrupamento

Análise Multivariada de Dados

	Variáveis					
Unidades Amostrais	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}

Objetivos:

Análise no $\Re^{nxn}$
Redução de dimensionalidade das unidades amostrais!

- Formação de grupos de unidades amostrais $\Rightarrow$ agrupamento de observações $\Rightarrow$ grupos homegêneos internamente e heterogêneos externamente
- 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

D: Euclidiana

Matrizes de Distância

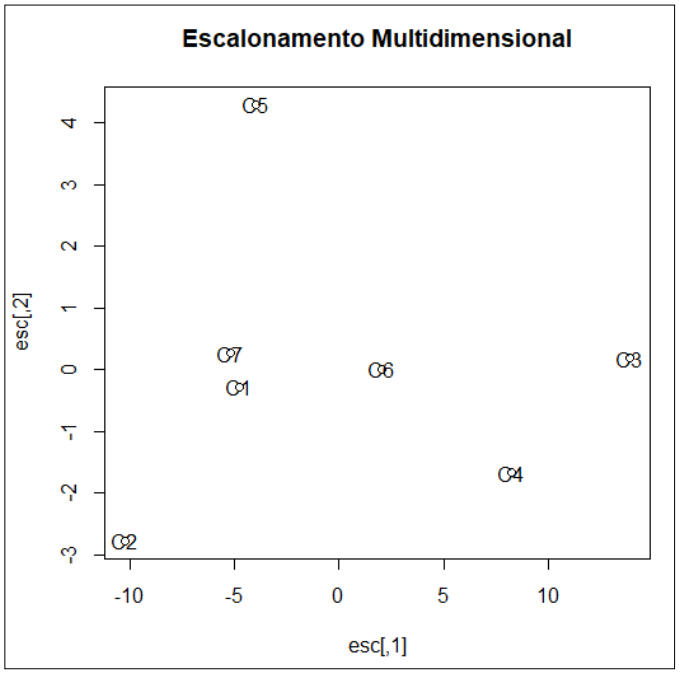
	1	2	3	4	5	6	7
1	0	6.21	18.70	13.13	4.83	7.43	2.03
2	6.01	0	24.34	18.55	9.44	12.94	6.62
3	18.67	24.31	0	5.99	18.38	12.50	19.20
4	13.10	18.52	5.94	0	13.64	7.26	13.78
5	4.64	9.44	18.35	13.62	0	7.98	5.09
6	6.78	12.55	11.89	6.48	7.36	0	8.67
7	0.70	5.87	19.11	13.61	4.21	7.22	0

$= \hat{D}$

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

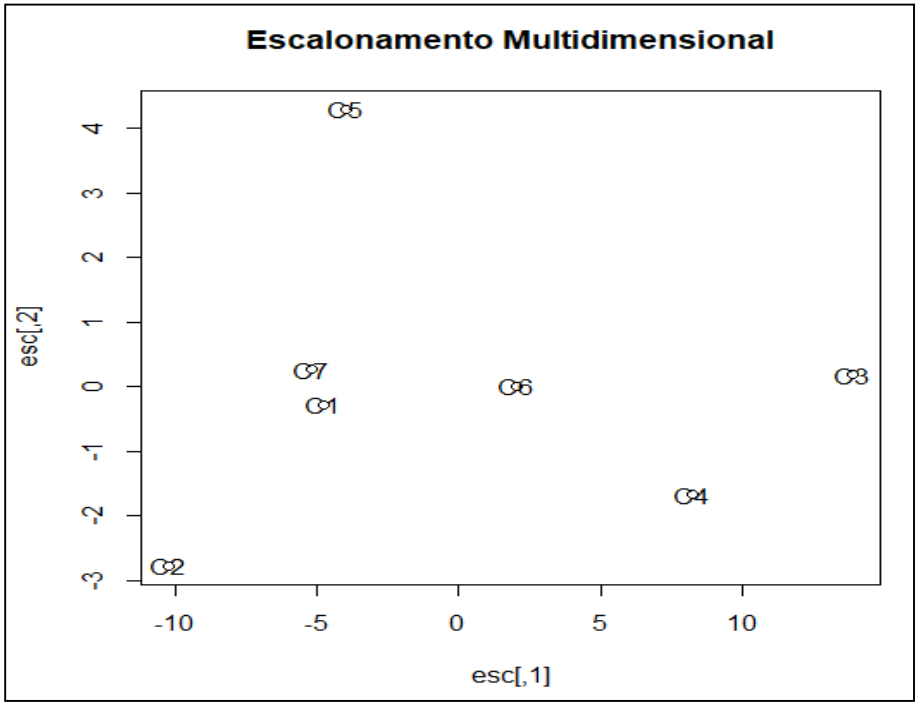


Representação das observações em $\Re^2$: soluções equivalents de **Componentes Principais** e **Escalonamento Multidimensional**

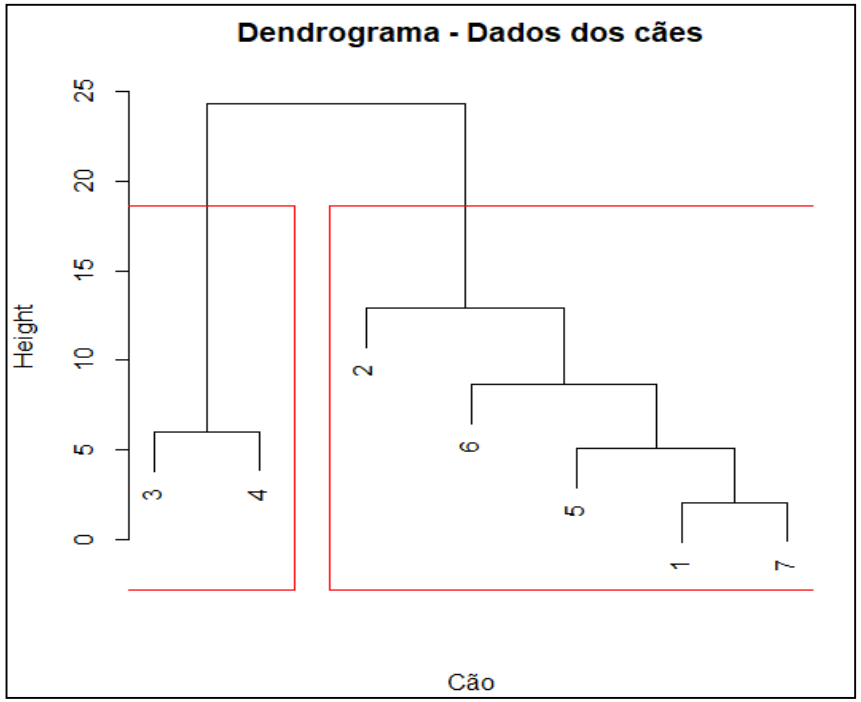
Análise de Agrupamentos

Escalonamento Multidimensional

Coordenadas Principais



Análise de Agrupamento



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 v*ariáveis originais ou padronizadas)*.
- **Definicão 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 Similaridade (Correlação): quanto maior o valor, maior a semelhança entre os objetos.

- 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

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

Distância de Manhatan (quarteirão)

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

Distância de Minkowsky (mais geral)

Análise de Agrupamentos

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 distância

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

- Nem sempre é possível adotar a correlação como medida de parecença entre "unidades amostrais" (Ex. Considere o gráfico de dispersão de duas obs x e y)

- O coeficiente de correlação *r* é comumente usado como medida de "parecença" entre variáveis (e não entre observações, isto é, entre unidades amostrais)

- A correlação valoriza padrões de forma (tendências) 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

$$d(G_1, G_2) = \max_{i \in G_1, k \in G_2} d_{ik}$$

$\Rightarrow$ 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

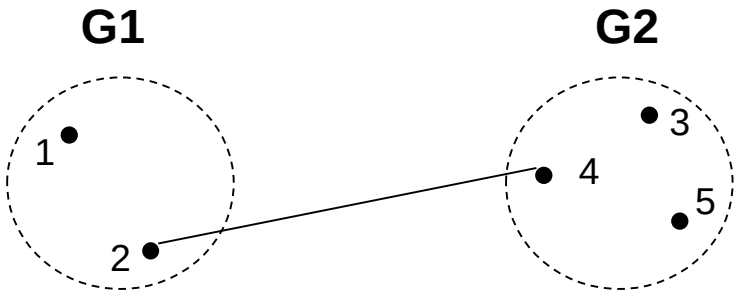
$$d(G_1, G_2) = \min_{i \in G_1, k \in G_2} d_{ik}$$

$\Rightarrow$ 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

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

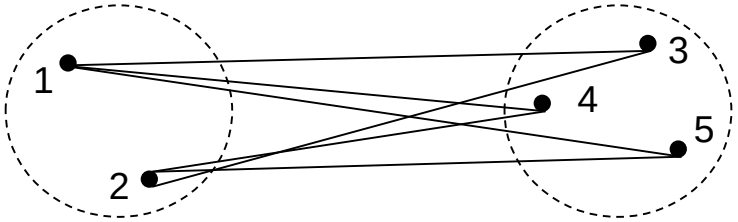
Algoritmos Hierárquicos



Distância entre Grupos

Ligação Simples

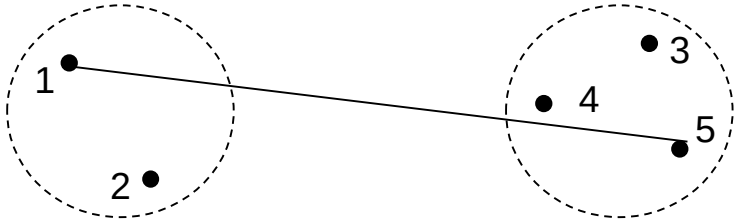
$$d(G_1, G_2) = d_{24}$$



Ligação Média

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

2x3

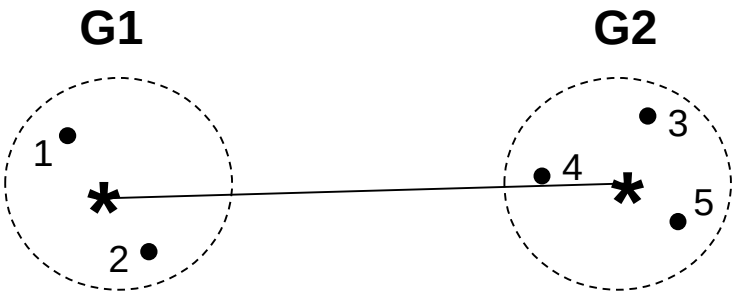


Ligação Completa

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

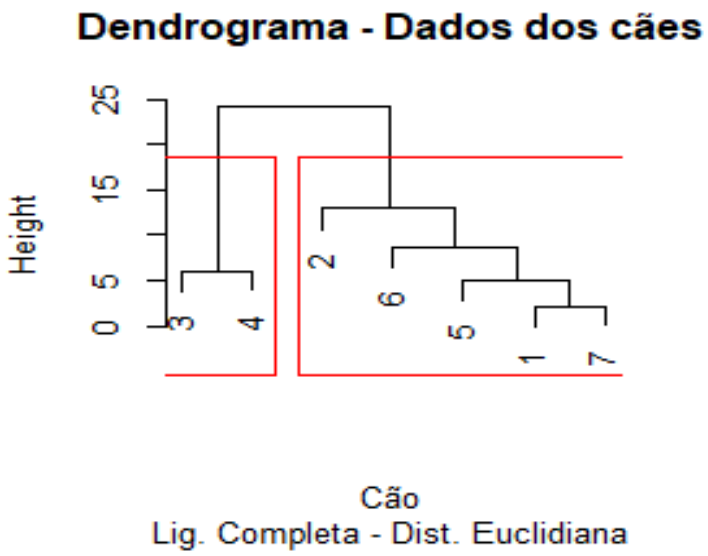
Algoritmos Hierárquicos

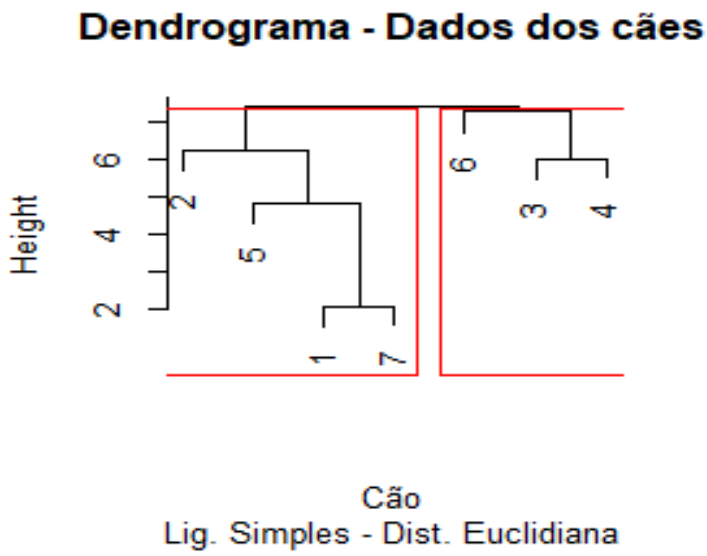
- **<u>Método da Centróide</u>**; este método define a coordenada de cada grupo como sendo a média das coordenadas de seus elementos. Uma vez obtida esta coordenada comum (denominada centróide) a distância entre os grupos G_1 e G_2 é dada pela distância entre as centróides.

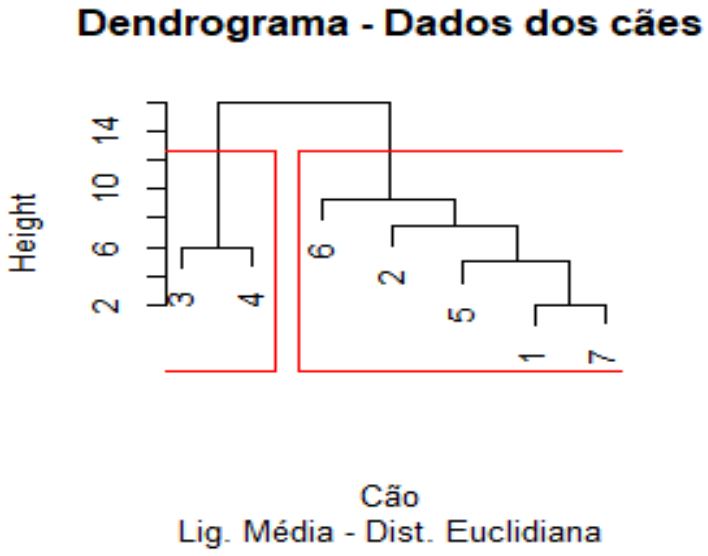


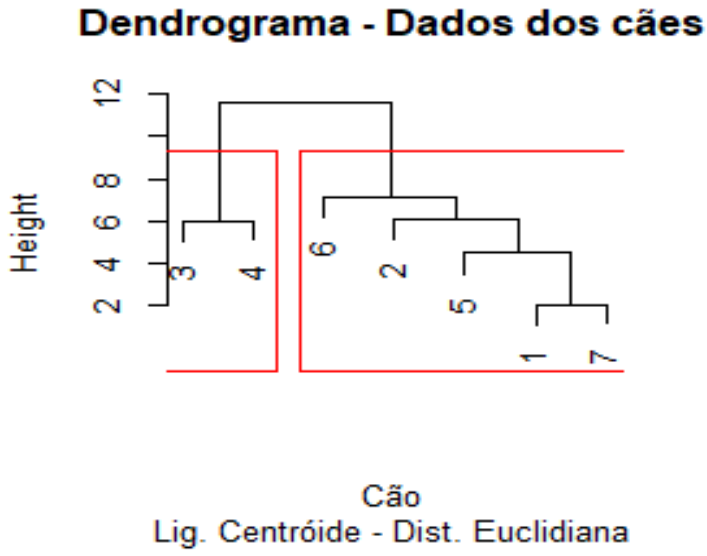
Análise de Agrupamento – Dados dos Cães

p=6

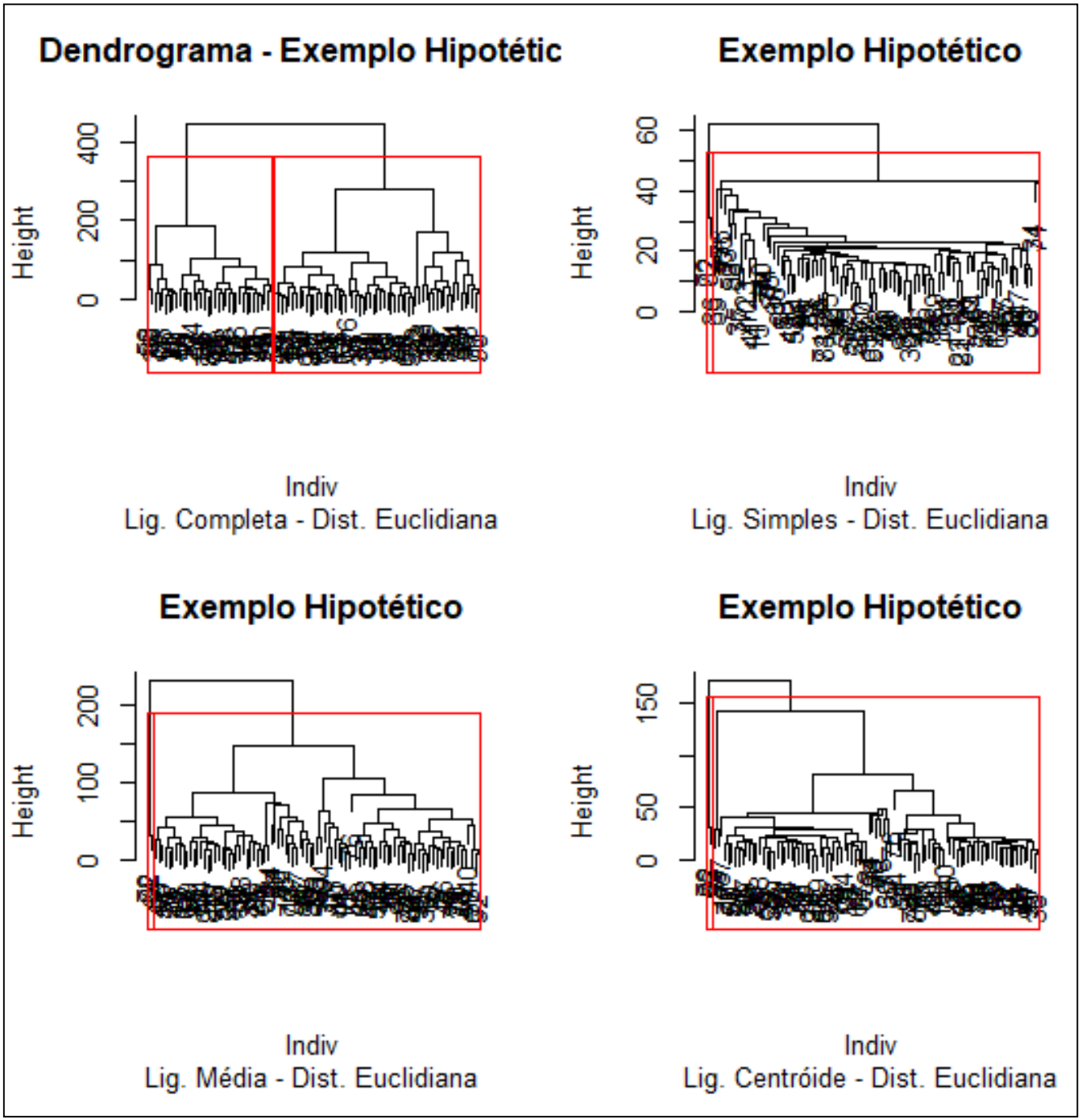








Estabelecer um ponto de corte (no eixo y) para a formação de grupos!
Neste caso, 2 grupos.



Dados:
Exemplo hipotético

Algoritmos Hierárquicos

- **<u>Método de Ward</u>**: 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).

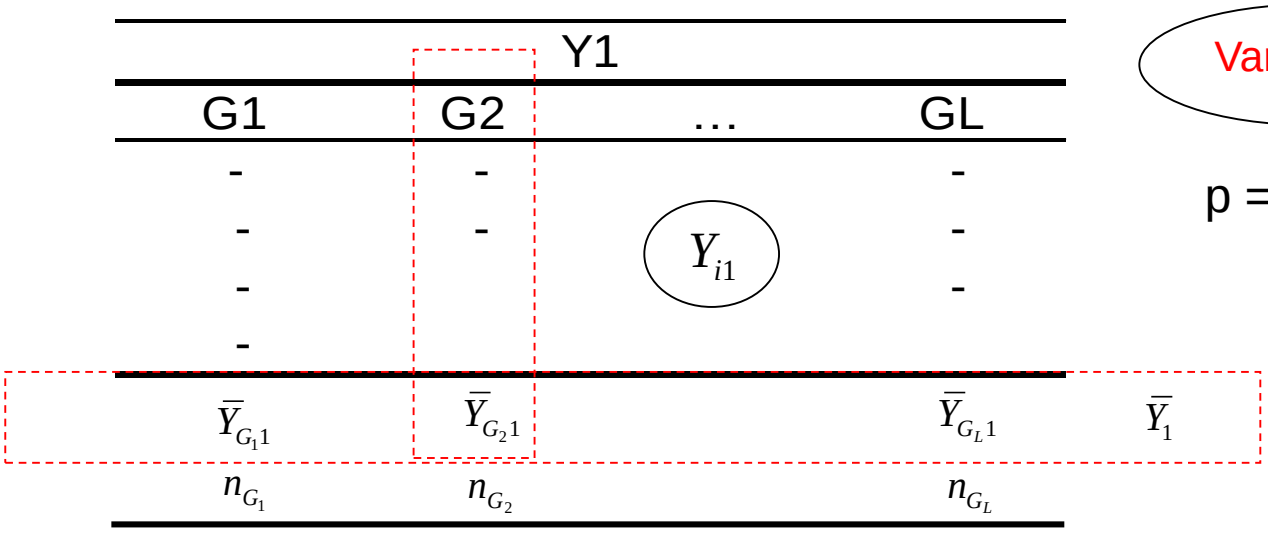
Considere a formação de L grupos de observações por meio de valores de uma única variável Y1.

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

$\overline{Y}_1$

Pense na ANOVA: como particionar a soma de quadrados total em componentes de variabilidade ENTRE e DENTRO de grupos?

Algoritmos Hierárquicos



Variável Y1

p =1 variável!

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

$$\sum_{l=1}^{L}\sum_{i\in G_l}\left(Y_{i1}-\bar{Y}_1\right)^2 = \sum_{l=1}^{L} n_{G_l}\left(\bar{Y}_{l1}-\bar{Y}_1\right)^2 + \sum_{l=1}^{L}\sum_{i\in G_l}\left(Y_{i1}-\bar{Y}_{l1}\right)^2$$

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

Algoritmos Hierárquicos

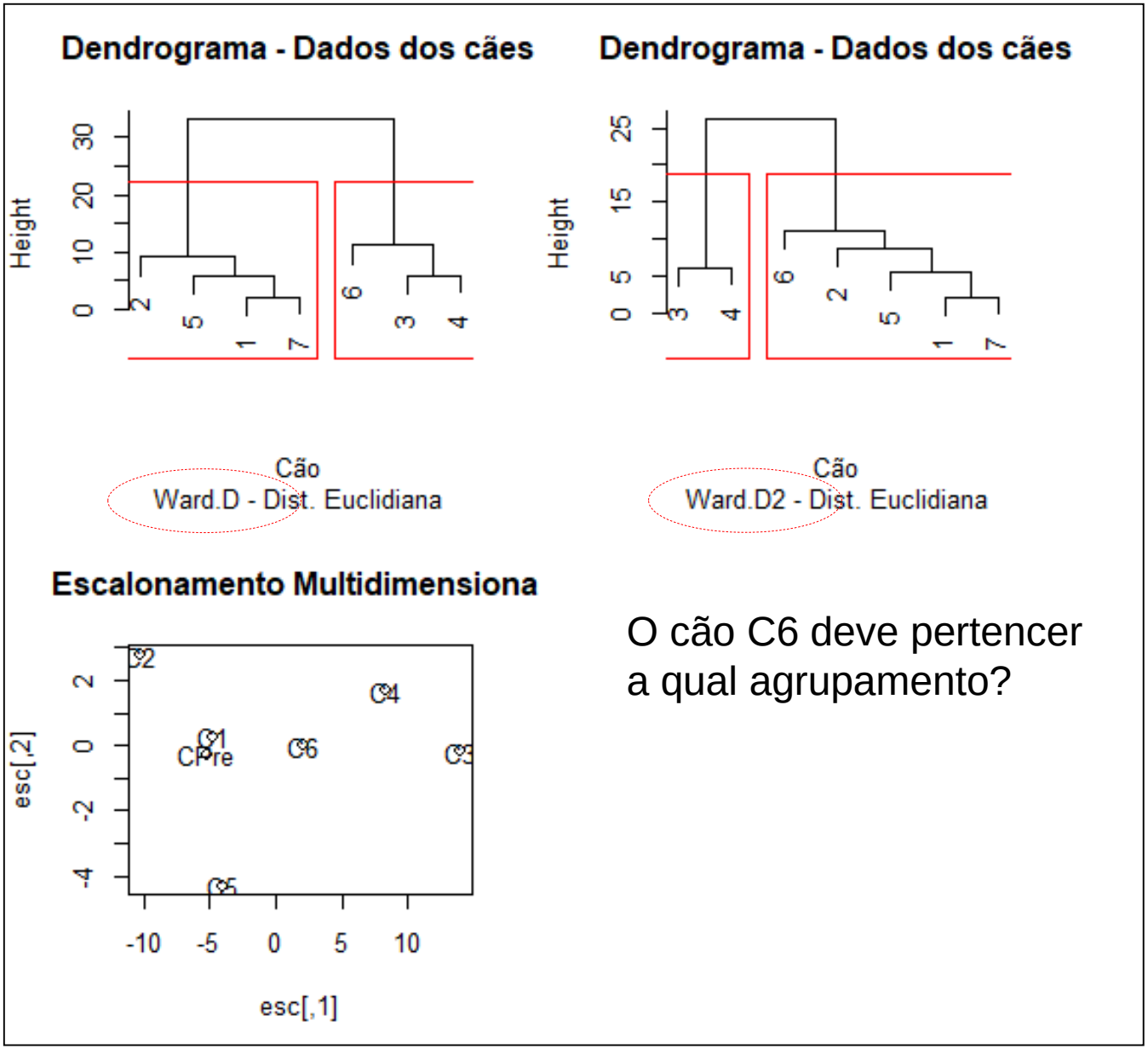
Método de Ward: Para considerar as p variáveis simultaneamente define-se a Soma de Quadrados (Dentro) da Partição como:

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

Procedimento:

- Passo 1: Para n observações, calcular SQDP para os possíveis (n-1) grupos distintos e selecionar o agrupamento com a menor SQDP
- Passo 2: Calcular SQDP para os possíveis (n-2) grupos distintos (fixada a união de 2 obs 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.

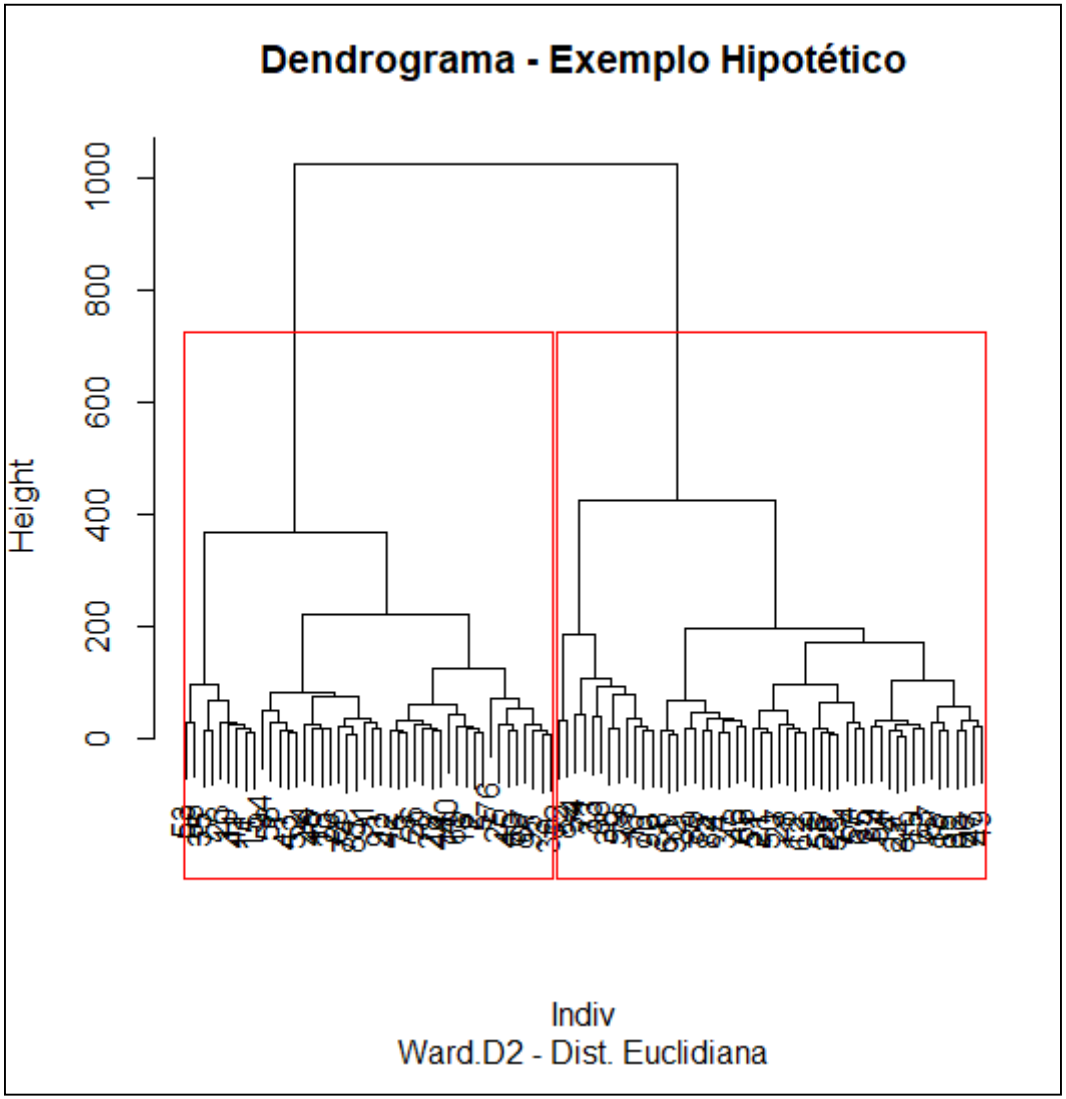
Análise de Agrupamentos – Dados dos Cães



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!



Dados:
Exemplo hipotético com 3 grupos genotípicos

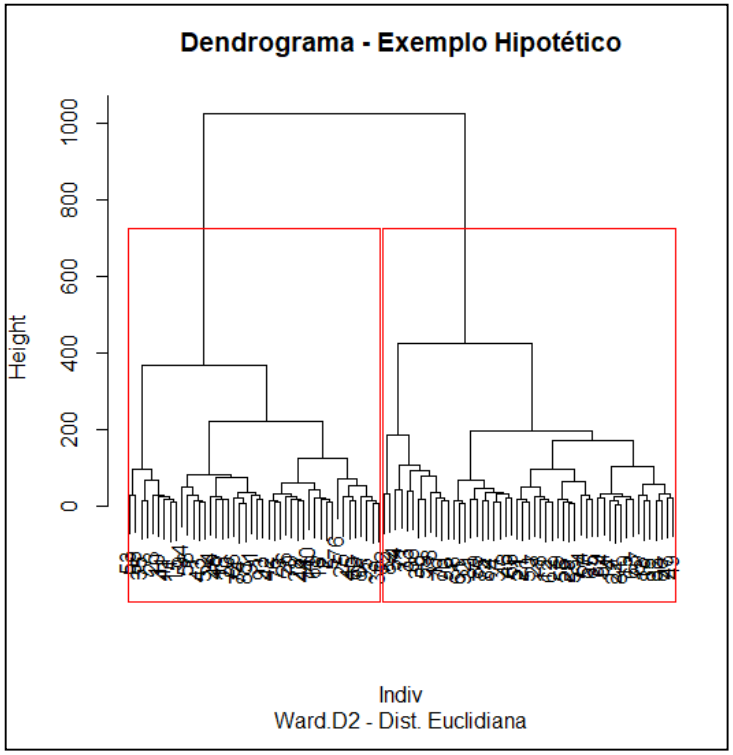
Caracterização dos grupos:

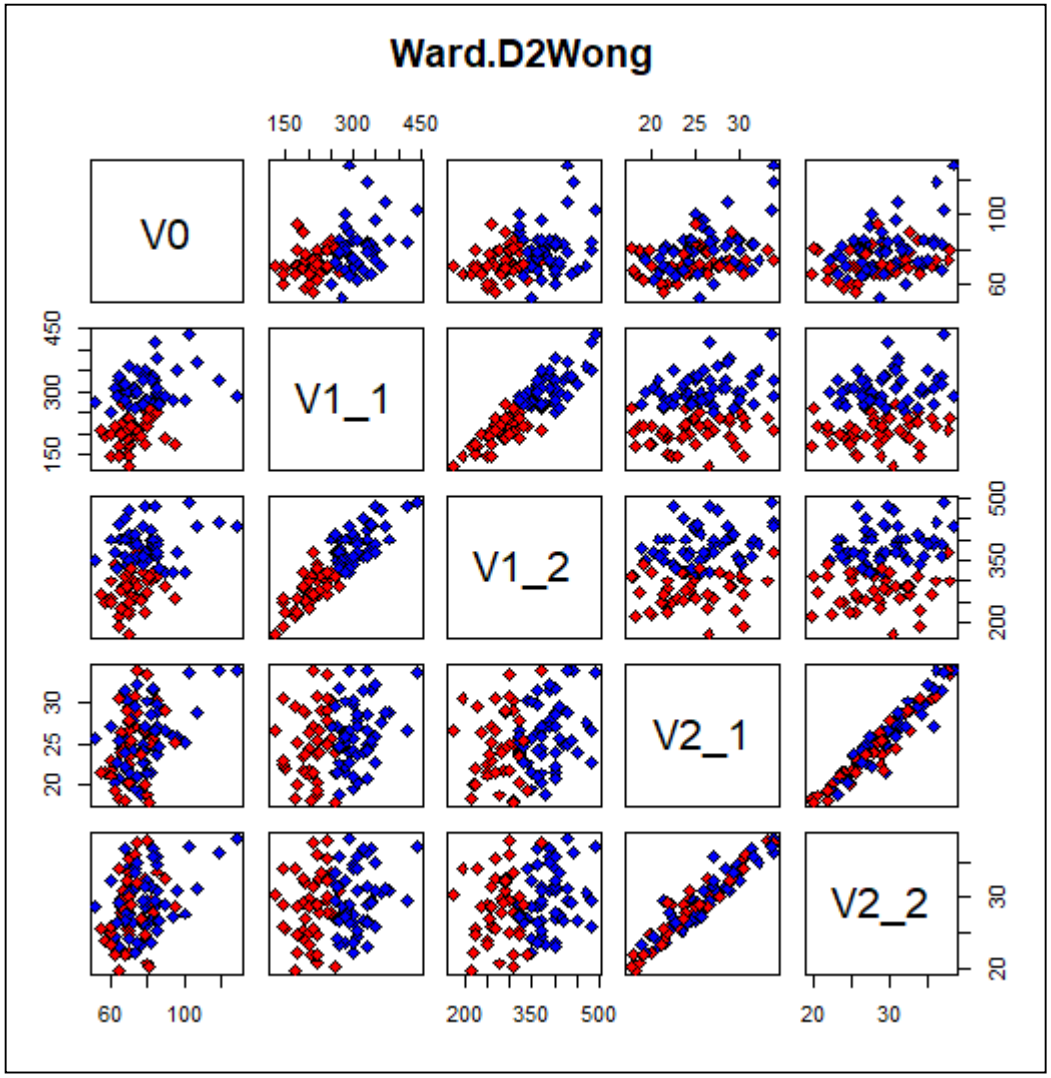
Gen	G1	G2
1	8	9
2	23	27
3	13	15

Grupos:

```
 [1] 1 1 1 2 1 1 1 2 2 1 1 2 2 2 1 1 2 1 2 1 2 2 1 1 1 1 2 1 2 1 2 2 2 2 1
[36] 2 2 2 2 1 1 1 1 1 1 1 2 2 2 2 2 1 1 2 2 2 2 1 2 1 2 2 2 2 2 2 1 2 1 1
[71] 2 2 1 1 2 1 1 2 2 2 2 2 1 2 2 1 2 1 2 2 1 2 1 1 1
```

Dados: Exemplo Hipotético





Caracterização dos grupos:

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 funçao 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 recalcular o respectivo centróide.
- Realocar a observação removida em algum dos grupos, de forma a **minimizar** a quantidade D. Recalcular 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

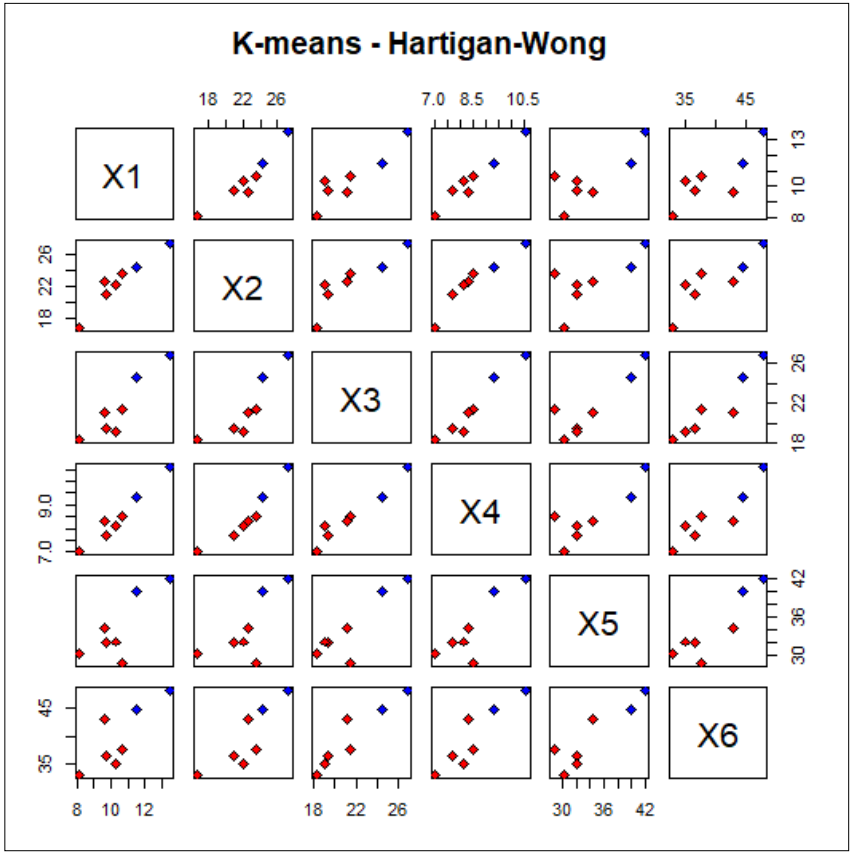
Dados dos Cães

```
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 = R^2 = 71.9%

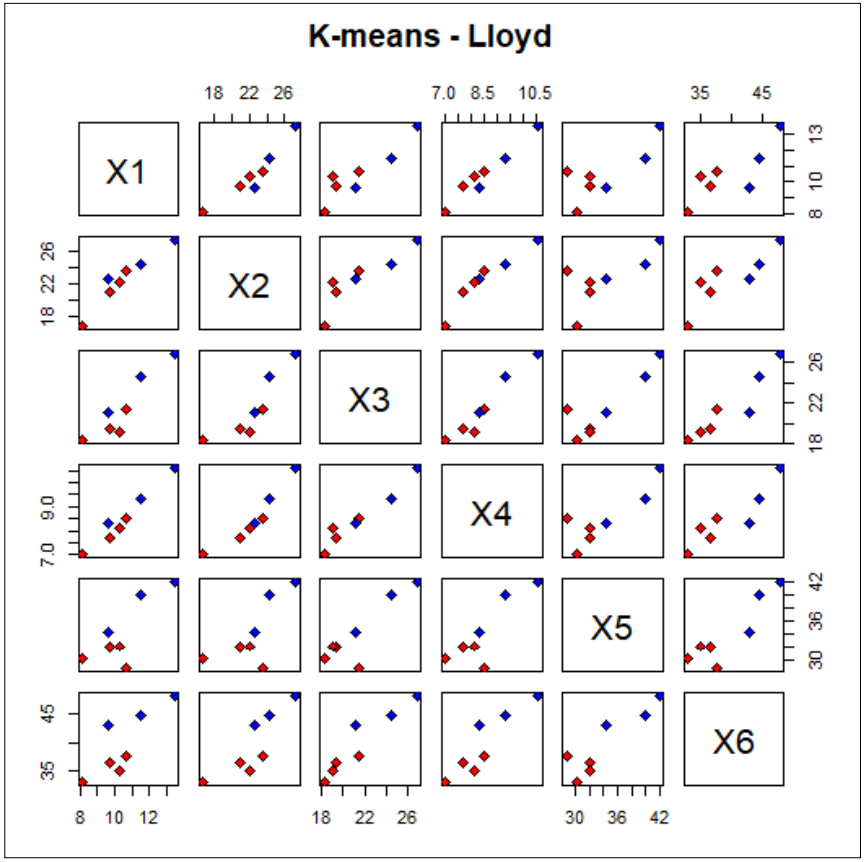
```
Algoritmo "Lloyd":
C1 C2 C3 C4 C5 C6 C7
 1  1  2  2  1  2  1   SQE/SQTotal = 71.4%
```

Método das Partições: K-Médias



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

Análise de Agrupamento

Medidas biométricas (mm) de Pardais fêmea

(Manly, 2005; Hermon Bumps, 1898).

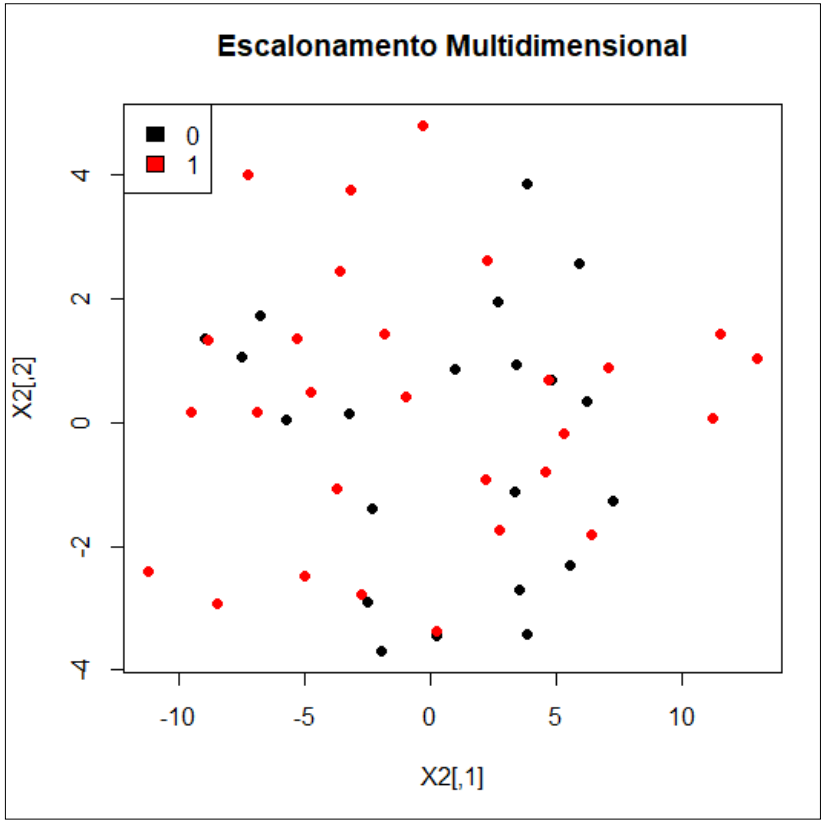
Pardal	Sobrev.	X1	X2	X3	X4	X5
1	S	156	245	31.6	18.5	20.5
…	…					
21	S	159	236	31.5	18.0	21.5
22	N	155	240	31.4	18.0	20.7
…	…					
49	N	164	248	32.3	18.8	20.9

Dados dos Pardais:

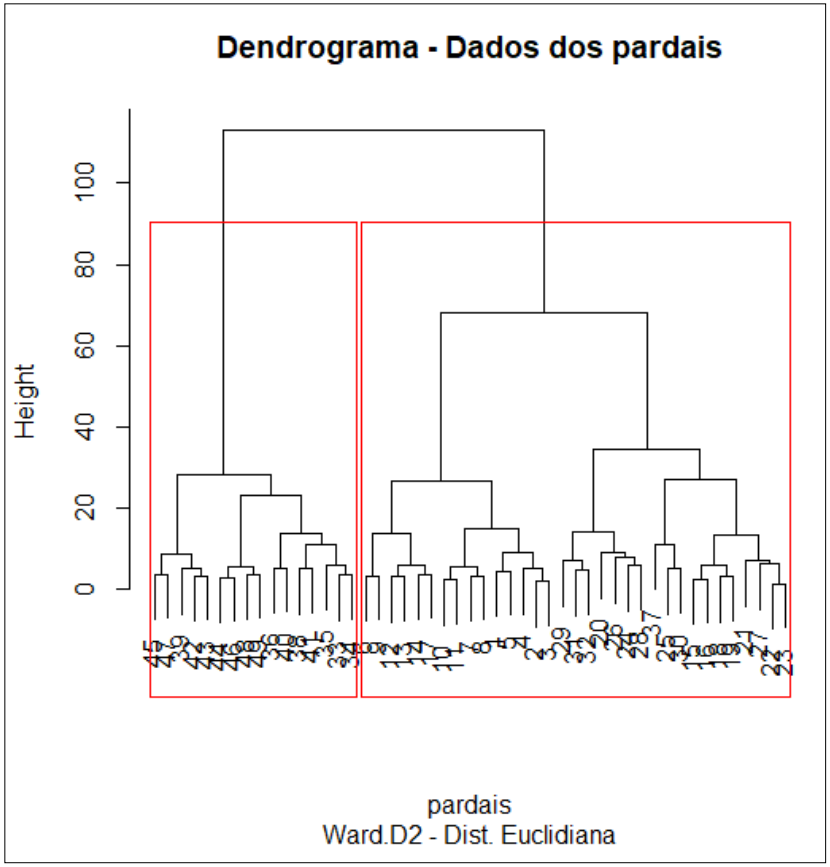
-**Análise de CoP**: permite a visualização dos dados nos grupos Sobreviventes (=0) e Não Sobreviventes (=1)

-**Análise Discriminante**: Como os grupos de pardais Sobreviventes (=0) e Não Sobreviventes (=1) podem ser preditos?

- **Análise de Agrupamentos** (k=2): As variáveis biométricas (X1 a X5) permitem a classificação dos pardais em Sobreviventes e Não Sobreviventes?



Análise de Agrupamento




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

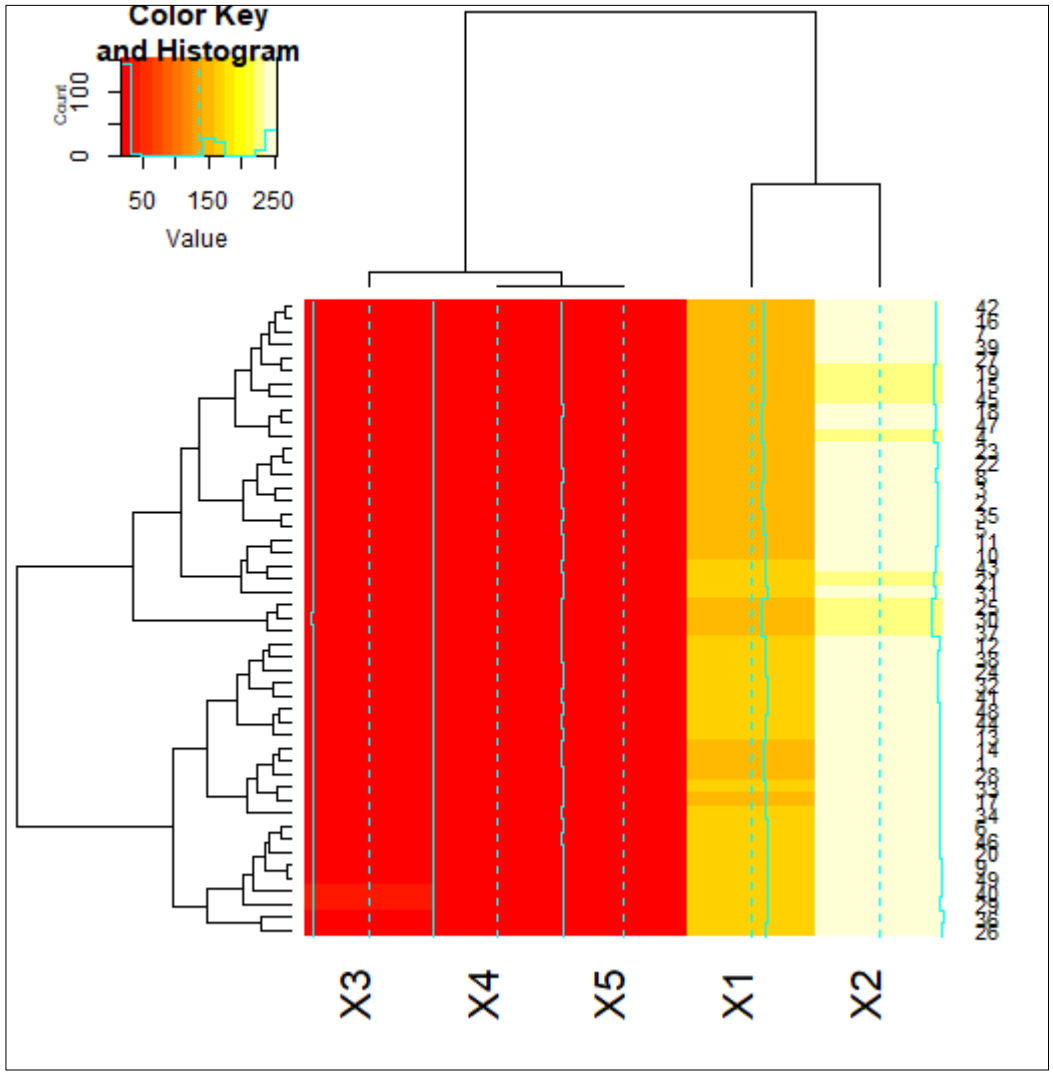
Algoritmo k-Means Lloyd

```
      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
```

Análise de Agrupamento
Aplicação: Heatmap



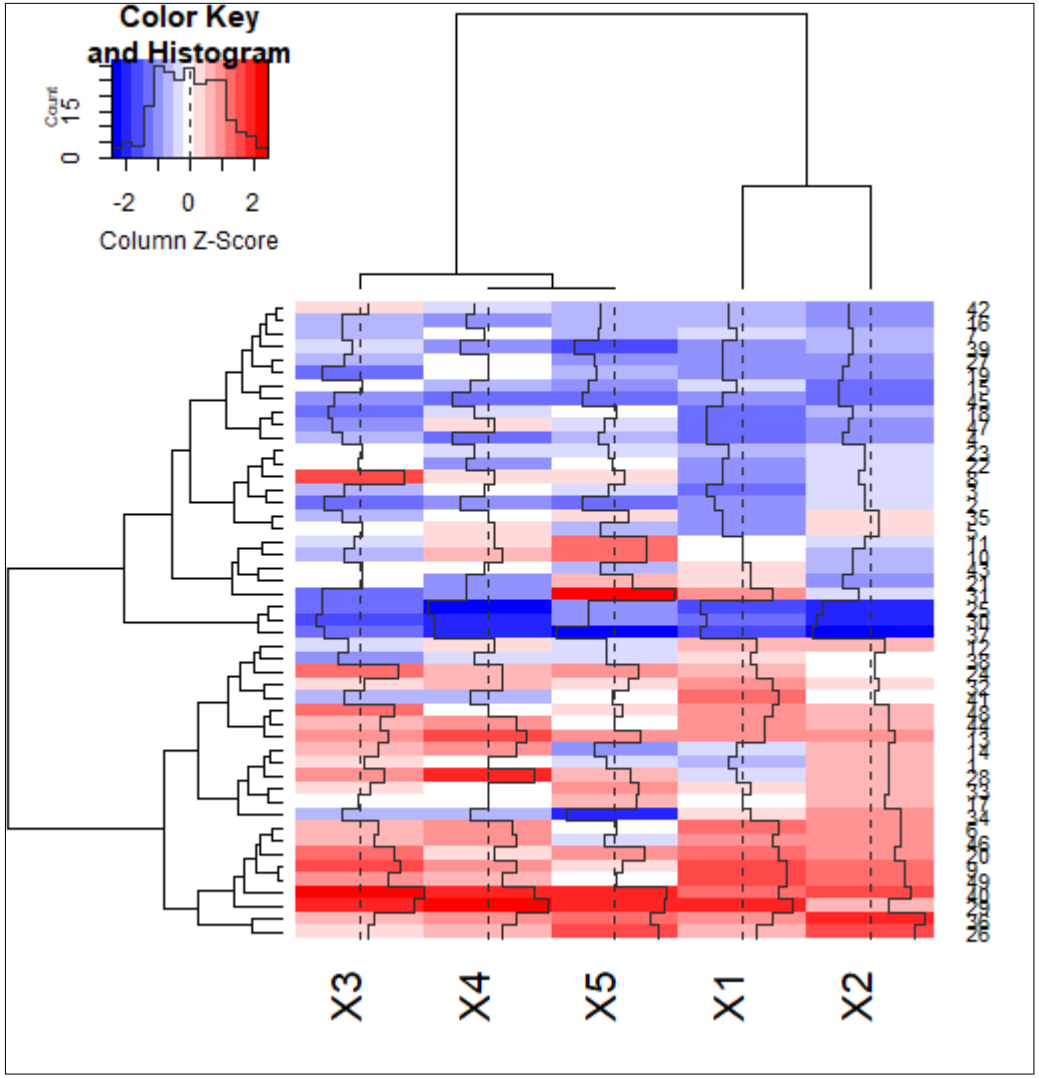
Dados dos Pardais
Representação simultânea das unidades amostrais e das variáveis

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

As distâncias Eucludianas definem as cores.

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

Análise de Agrupamento
Aplicação: Heatmap



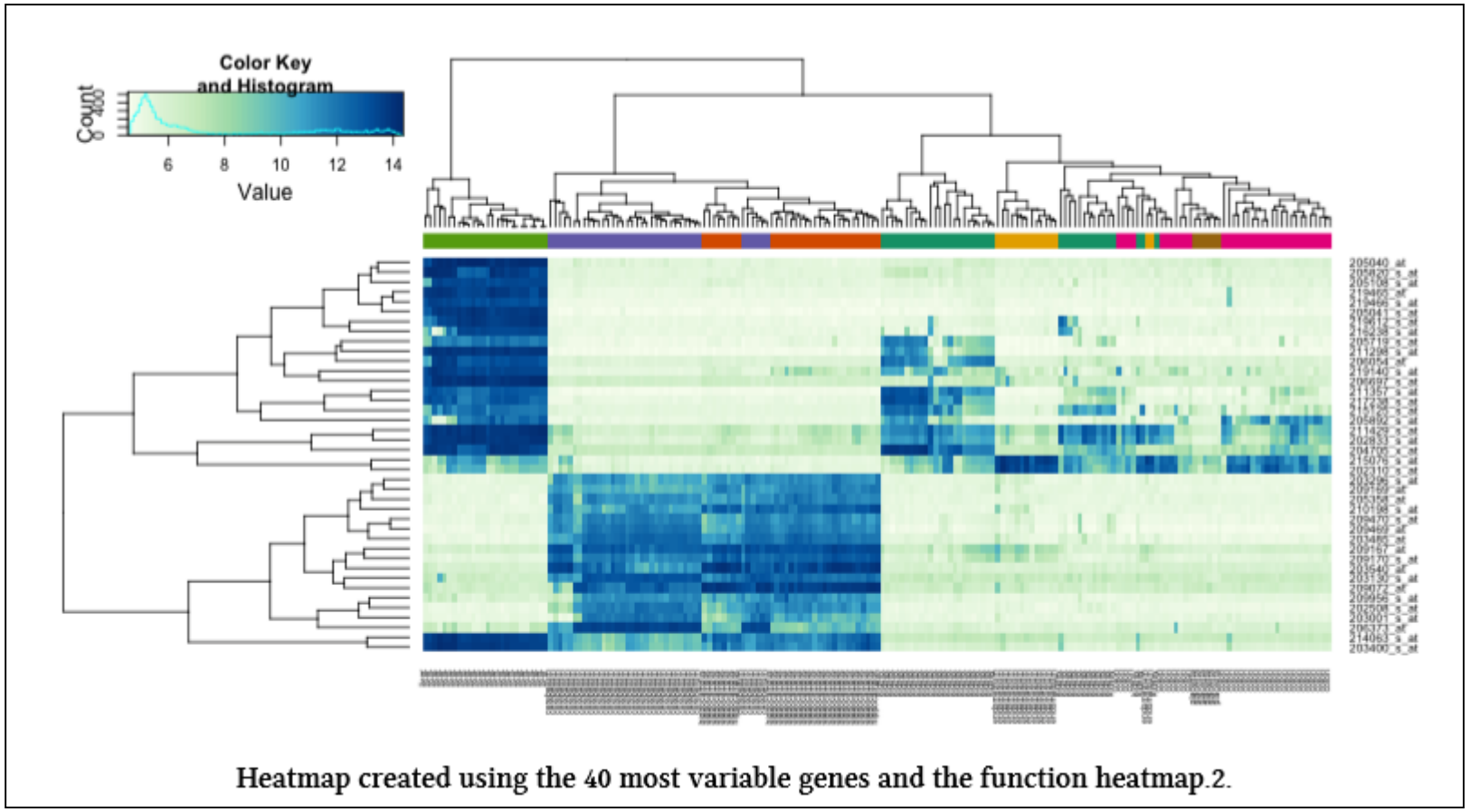
Dados dos Pardais

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, o primeiro é caracterizado pelos maiores valores do escorez 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)