Métodos de agrupamento

Baseado em: “An Introduction to Statistical Learning with R - 2º edition”

Por: Pedro de Araújo Ribeiro

Considerações:

- Biblioteca usada:
  - Datasets (nativa do R)
- Conteúdos abordados:
  - Agrupamento por K-média
  - Agrupamento hierarquico
  - Considerações gerais de agrupamentos
  - Prática

Proposta:

Certas análises podem demandar um estudo estatístico não contemplado por regressões numéricas ou algoritmos de classificação, sendo assim abordaremos métodos de agrupamento de observações.

Ao final, faremos um estudo de um conjunto de dados referente aos hábitos alimentares europeus durante a Guerra Fria.

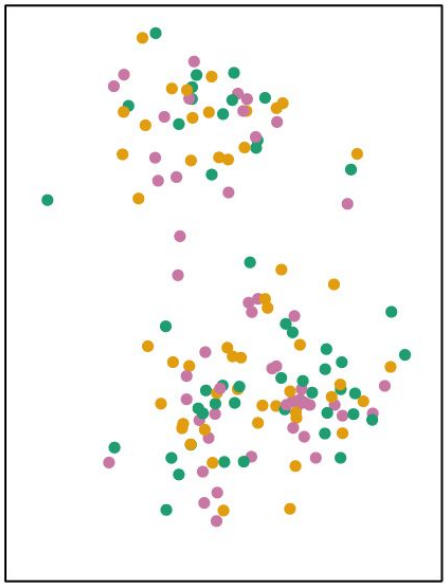
Considerações iniciais:

Ao longo desta apresentação lidaremos bastante com conceitos de distância e semelhança, para tal estaremos trabalhando com variáveis numéricas e (usualmente) normalizadas.

A função de normalização fica a critério do leitor e não será contemplada aqui, porém normalização a partir do desvio padrão é a mais comum.

K-médias:

Consiste de estabelecer previamente *K* aglomerados e preenche-los com observações de tal modo que a variação entre as observações dentro de cada aglomerado seja a mínima possível, nos permitindo intuir relações e tendências entre tais observações.



O processo começa com a atribuição aleatória de observações a grupos e a partir daí aplicamos um algoritmo que reduzirá progressivamente a diferença interna dos aglomerados.

Utilizaremos um exemplo com 2 variáveis e *K*=3 para facilitar a explicação, porém o método é válido para qualquer quantidade de variáveis numéricas.

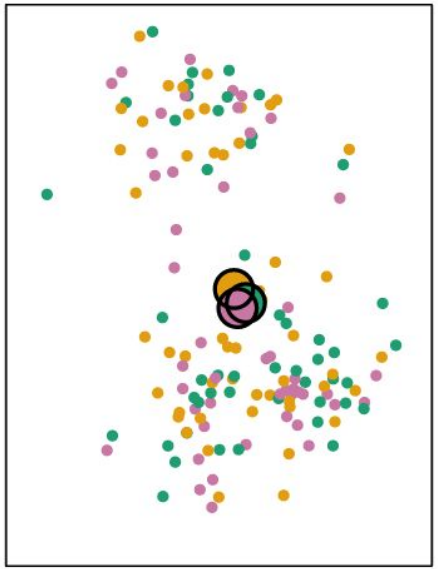
K-médias:

Tendo realizado as atribuições, calculamos o centróide dos *K* aglomerados a partir da soma dos valores de uma variável *x* e dividindo pela quantidade de observações, repetimos para todas as variáveis para obter as coordenadas do centróide.

O centróide pode ser descrito pela fórmula abaixo para *n* observações da variável *x*:

$$C = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

No exemplo os centróides acabaram muito próximos devido a aleatoriedade da atribuição inicial, porém este não será sempre o caso.

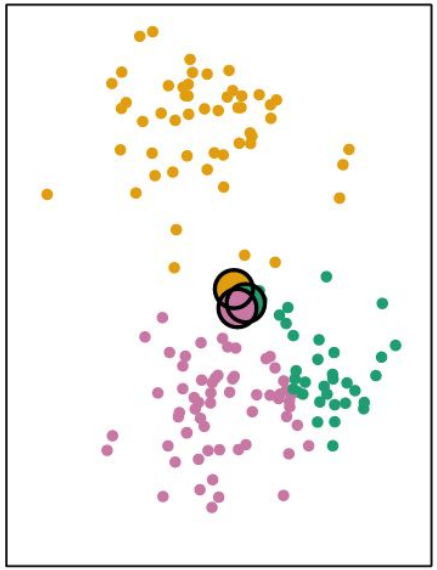


K-médias:

Agora faremos a reatribuição dos grupos, onde cada observação será associada ao centróide mais próximo. Utilizaremos como medida de proximidade a distancia euclidiana, descrita pela seguinte fórmula:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}.$$

Novamente, estamos trabalhando com apenas 2 variáveis e portanto 2 dimensões, porém esta fórmula é válida para *n* variáveis.

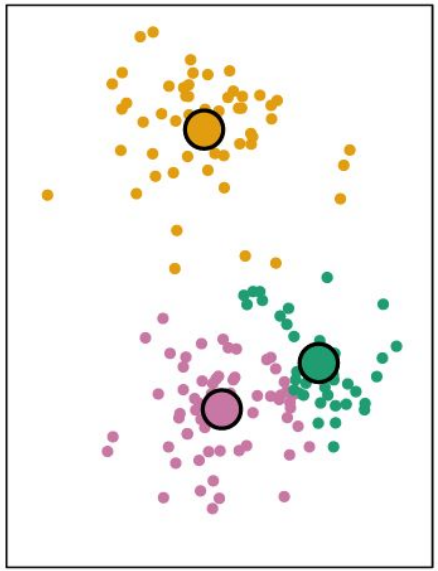


K-médias:

Em seguida recalculamos o centróide, neste momento devemos também calcular a taxa de variação dentro de cada grupo. Esse cálculo é dado pela média da soma dos quadrados das distâncias entre todas as observações.

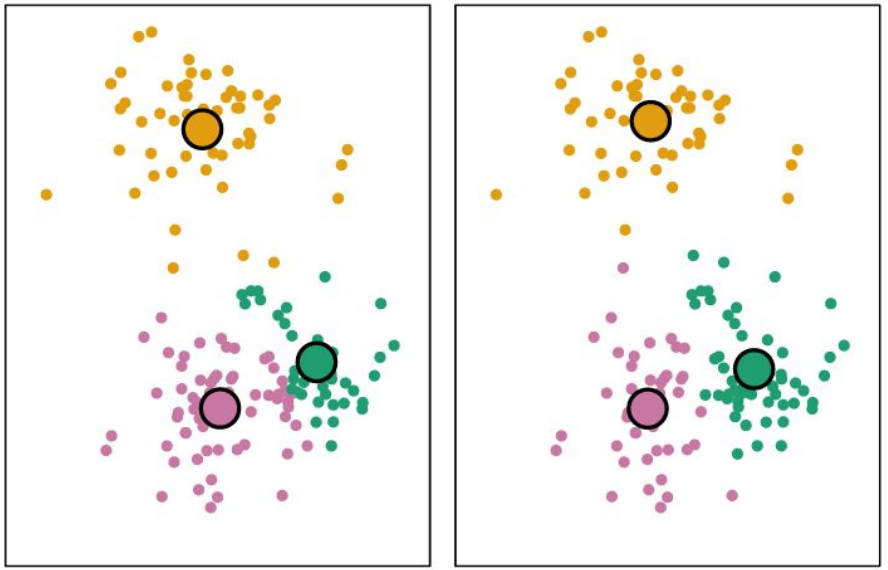
$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2,$$

Esta fórmula representa a variação de apenas um dos grupos, porém nosso objetivo é minimizar a variação geral, obtida a partir da soma das variações individuais.



K-médias:

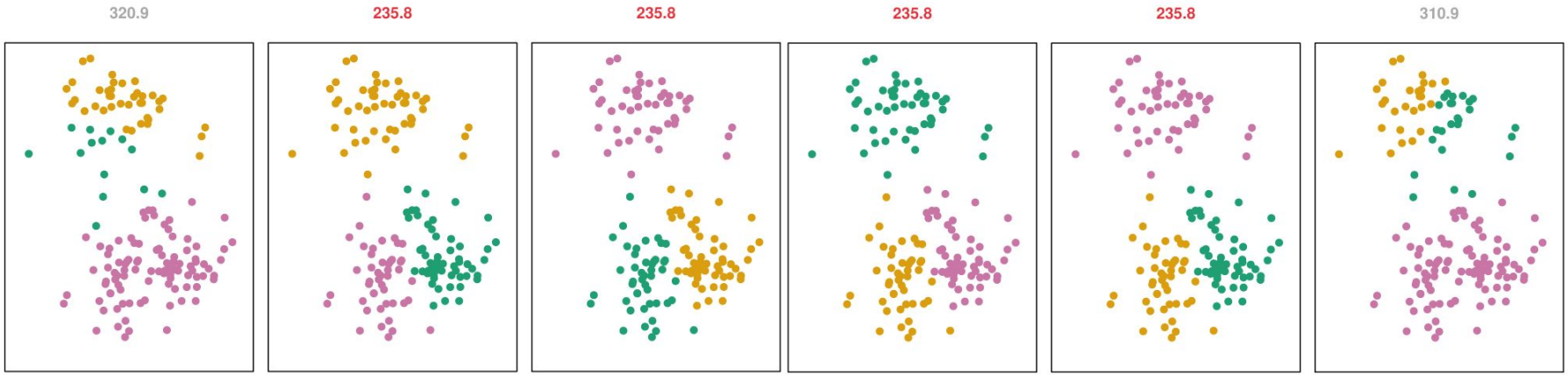
Tendo recalculado os centróides podemos redistribuir os grupos baseado novamente na distância mais próxima do centro e ao final calculamos mais uma vez a taxa de variação geral. Esse processo se repete até que a taxa pare de mudar com iterações seguintes.



Desvantagens do K-médias: Aleatoriedade

Devido a distribuição aleatória inicial, o restante do processo será inseparável dos primeiros centróides calculados, fazendo com que diferentes aplicações do método levem a resultados diferentes.

Exemplo:



Desvantagens do K-médias: Falta de nuance

A menos que um pesquisador revise manualmente os cálculos de distancias euclidianas internas na etapa do cálculo da variação, é impossível quantificar o quão “parecidas” duas variáveis que estão no mesmo grupo realmente são.

Comparado com métodos hierarquicos que veremos em seguida, esta perspectiva mais generalista leva a análises e conclusões com menos nuance.

Desvantagens do K-médias: Suposições

Possivelmente a maior desvantagem do método e a necessidade de assumir previamente um K, o que é uma decisão muito subjetiva e dependente da natureza das informações a serem analisadas.

Um K muito grande pode levar a observações agrupadas que na prática não são tão semelhantes assim, enquanto um K muito pequeno levaria a muitos grupos parecidos entre si e que se analisados separadamente não levariam a conclusão nenhuma.

Agrupamento hierárquico:

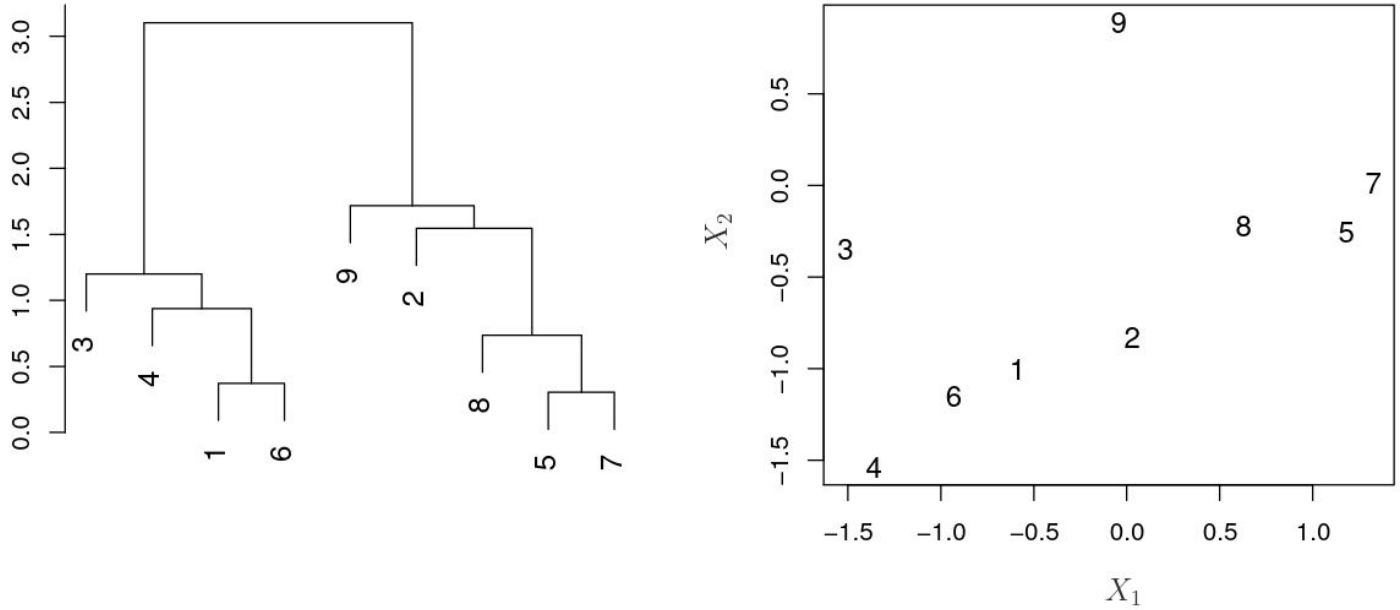
Consiste de tomar as *n* observações iniciais, para cada uma calcular a distância euclidiana (ou outra medida de semelhança) com as demais e, dentre todas as distâncias calculadas, tomar a menor de todas e transformá-la em um agrupamento.

Em seguida, recalculamos a distância entre as demais observações e o aglomerado que fizemos. Depois pegamos novamente a menor distância de todas, seja ela entre observações, grupos, ou observações e grupos, e fazemos novamente um agrupamento.

O resultado será uma figura denominada dendograma, onde a altura da união é indicativa do quão próximas duas observações são e em qual ordem se conectam.

Agrupamento hierárquico:

O processo anterior é executado até que haja só um grupo.



Agrupamento hierárquico: Linkagem completa

Há quatro maneiras diferentes de computar e interpretar a distância entre grupos.

Linkagem completa:

Calculamos as distância entre todas as observações de dois grupos A e B e armazenamos a maior, repetimos isso para todos os pares de grupos. Dentre eles, o par de grupos com a menor distância armazenada será agrupado.

	a	b	c	d	e
a	0	17	21	31	23
b	17	0	30	34	21
c	21	30	0	28	39
d	31	34	28	0	43
e	23	21	39	43	0

g1 = (a,b)
dist(g1,c) = max(di(a,c),di(b,c) = max(21,30) = 30
dist(g1,d) = max(di(a,d),di(b,d) = max(31,34) = 34
dist(g1,e) = max(di(a,e),di(b,e) = max(23,21) = 23

portando g2 = (g1,e) = (e,a,b)

Agrupamento hierárquico: Linkagem única

Calculamos as distância entre todas as observações de dois grupos A e B e armazenamos a menor, repetimos isso para todos os pares de grupos. Dentre eles, o par de grupos com a menor distância armazenada será agrupado.

	a	b	c	d	e
a	0	17	21	31	23
b	17	0	30	34	21
c	21	30	0	28	39
d	31	34	28	0	43
e	23	21	39	43	0

g1 = (a,b)

ao final, g2 = (g1,e,c) = (e,c,a,b)

dist(g1,c) = min(di(a,c),di(b,c) = min(21,30) = 21
dist(g1,d) = min(di(a,d),di(b,d) = min(31,34) = 31
dist(g1,e) = min(di(a,e),di(b,e) = min(23,21) = 21

Agrupamento hierárquico: Linkagem média

Calculamos as distância entre todas as observações de dois grupos A e B e armazenamos a média entre elas, repetimos isso para todos os pares de grupos. Dentre eles, o par de grupos com a menor distância armazenada será agrupado.

	a	b	c	d	e
a	0	17	21	31	23
b	17	0	30	34	21
c	21	30	0	28	39
d	31	34	28	0	43
e	23	21	39	43	0

g1 = (a,b)

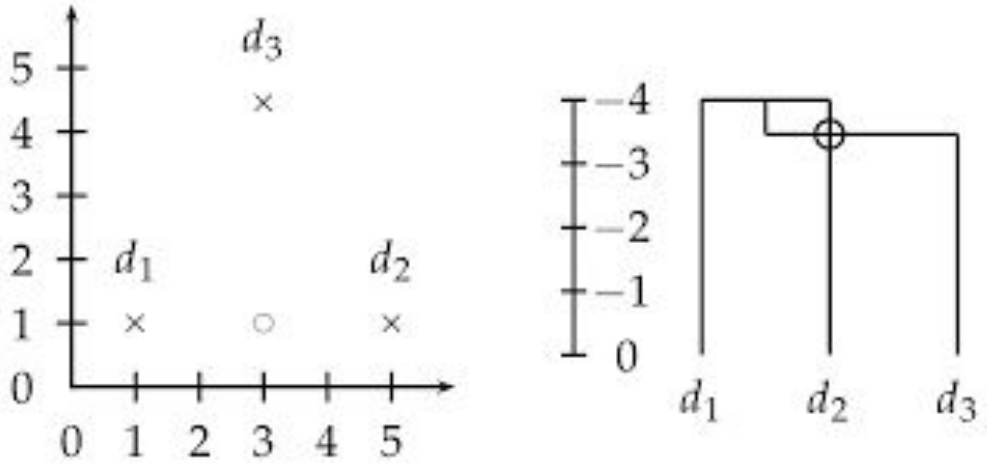
ao final, g2 = (g1,e) = (e,a,b)

dist(g1,c) = mean(di(a,c),di(b,c) = mean(21,30) = 25,5
dist(g1,d) = mean(di(a,d),di(b,d) = mean(31,34) = 32,5
dist(g1,e) = mean(di(a,e),di(b,e) = mean(23,21) = 22

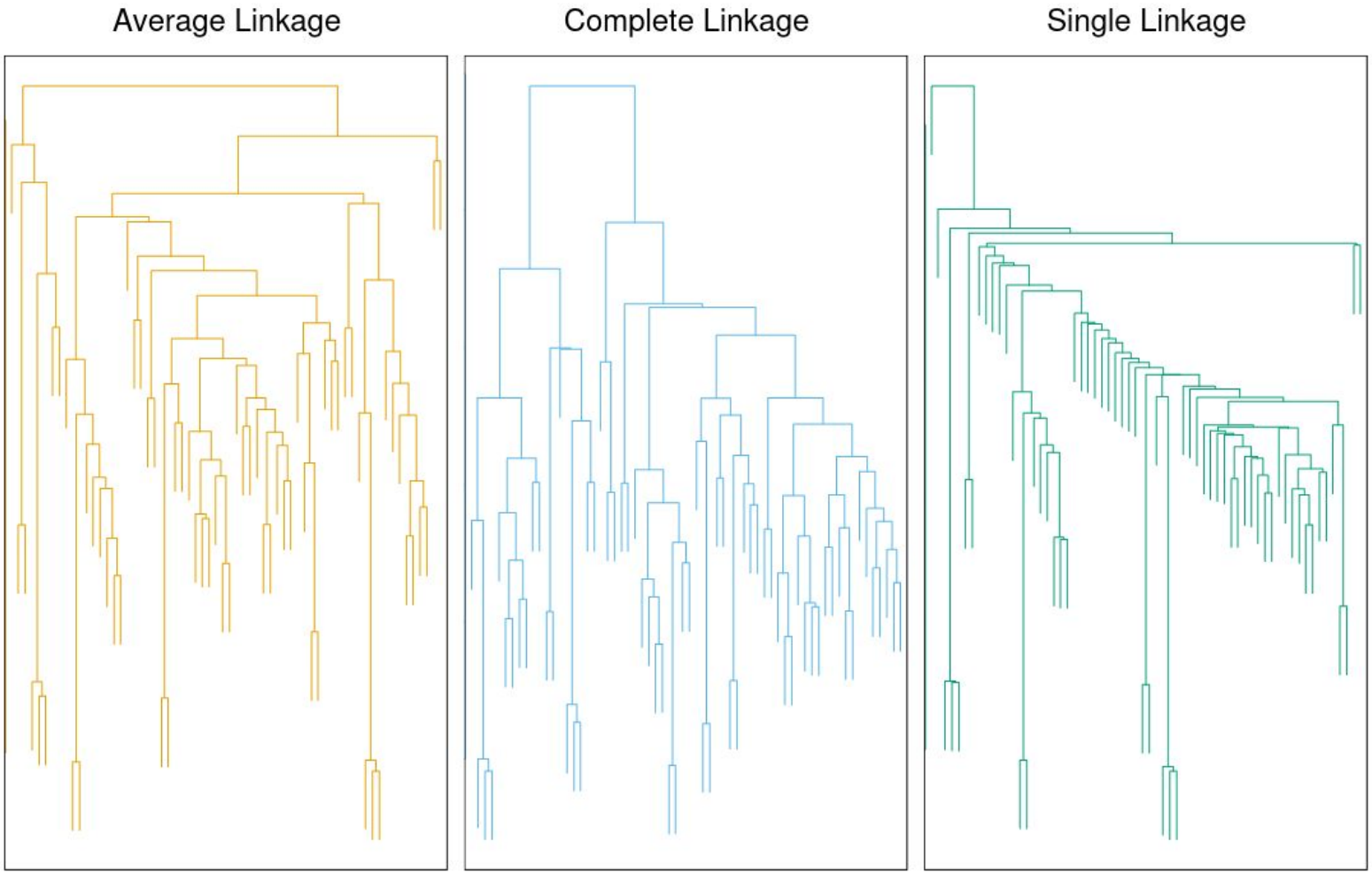
Agrupamento hierárquico: Linkagem por centróide

Calculamos o centróide de todos os grupos e analisamos as distâncias entre eles, escolhendo novamente a menor dentre todas (o cálculo do centróide se encontra nos slides anteriores).

Este é considerado o menos favorável de todos pois é o único que permite aberrações nos dendogramas, onde, devido a variância do centróide a cada iteração, é possível que um agrupamento “volte atrás”.



Agrupamento hierárquico:
Qual linkagem escolher?



Agrupamento hierárquico: Qual linkagem escolher

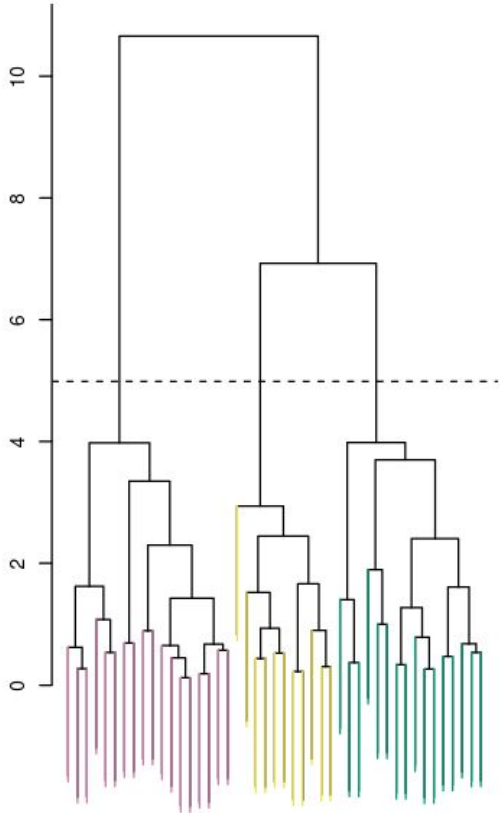
Esta é uma decisão extremamente subjetiva e dependente do problema a ser analisado, porém baseado no comportamento e definição dos métodos podemos observar que Média e Completa geram aglomerações mais balanceadas, enquanto a Única tende a agrupar “um a um”.

Contanto, vale ressaltar que o método do centróide é o pior e menos aplicado dentre os estudados.

Agrupamento hierárquico: Corte do dendrograma

Considerando que o agrupamento hierárquico termina quando todos as observações estão no mesmo grupo, devemos escolher um ponto de corte onde teremos X subgrupos que podemos analizar.

Essa decisão é *extremamente* subjetiva e depende dos preconceitos e expectativas estabelecidas antes do agrupamento assim como tendências observadas durante e após.

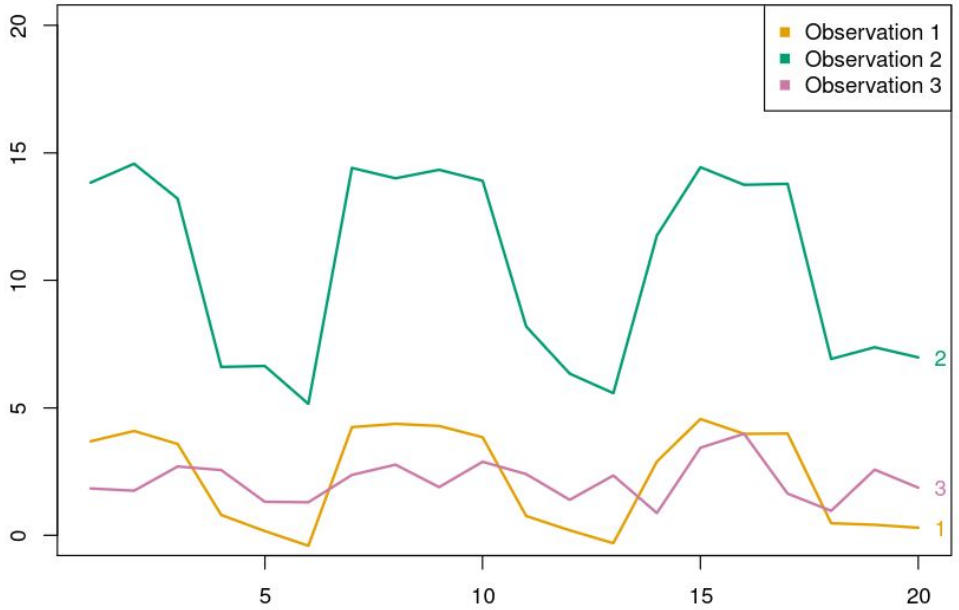


Agrupamentos: Medidas de semelhança

Até então nos referimos apenas a distância euclidiana como medida de semelhança entre observações e grupos, porém há outras métricas que levam em consideração diferentes critérios.

Uma delas é a "distância correlacional" que considera duas observações "próximas" se suas variáveis se comportam de maneira correlacional, independente de suas distâncias euclidianas.

(eixo x = índice da variável eixo y = valor da variável)



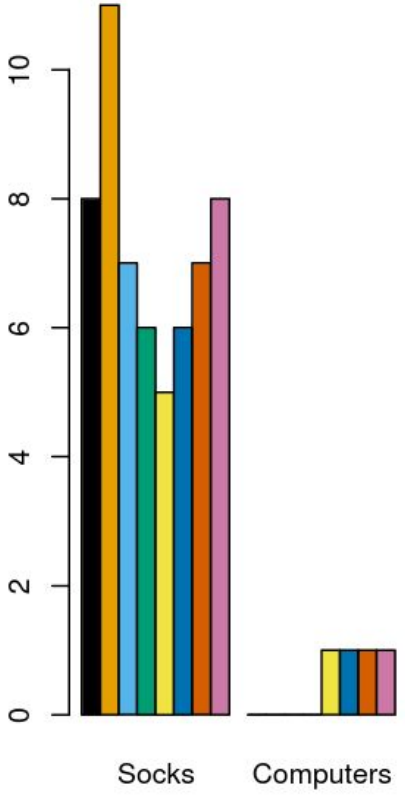
Agrupamentos: Subjetividade e análises

Uma parte considerável dos conceitos abordados exigem que o cientista ou analista tome decisões e assuma certas coisas, muitas vezes dependentes da situação problema e de seu próprio bom-senso e conhecimento acerca do assunto.

Abordaremos um exemplo onde um vendedor online vende apenas meias e computadores e queremos agrupar seus 8 clientes.

Agrupamentos: Subjetividade e análises - Exemplo

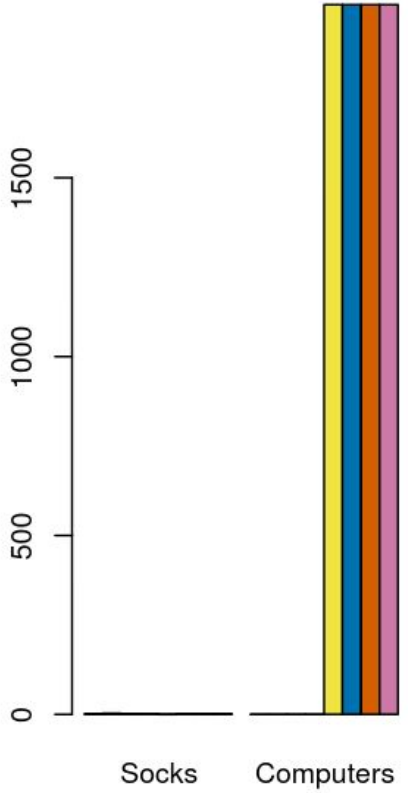
Neste primeiro gráfico temos a quantidade de itens comprados, sabendo que meias são consideravelmente mais baratas que computadores, além de serem compradas em maior quantidade, um agrupamento baseado na distância euclidiana utilizando as quantidades seria extremamente tendencioso as meias.



Agrupamentos: Subjetividade e análises - Exemplo

Agora temos os mesmos dados porém baseados na quantidade de dinheiro gasta pelos cliente. Aqui vemos que, como no exemplo anterior, uma das variáveis está dominando a análise.

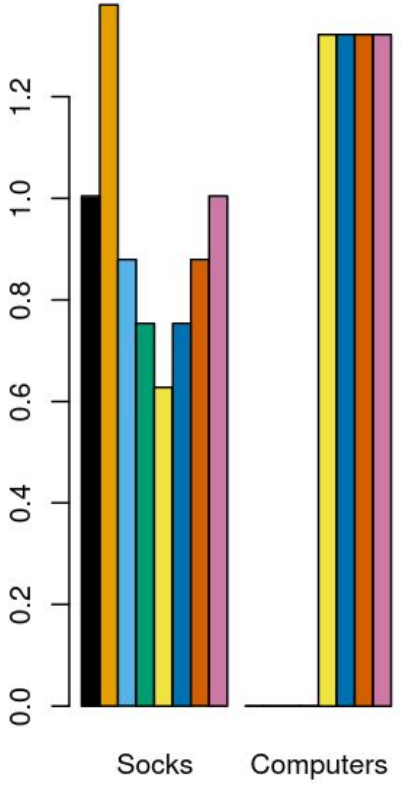
Neste caso, o agrupamento seria basicamente feito em função dos computadores.



Agrupamentos: Subjetividade e análises - Exemplo

Por fim, temos um gráfico onde as variáveis de quantidade foram normalizadas em função do desvio padrão. Aqui ambas são dadas mais relevância porém nenhuma domina a outra.

Via de regra será sempre mais proveitoso normalizar as variáveis, porém é importante analisar bem a situação em estudo para ter certeza.



Agrupamentos: Considerações finais

Agrupamentos são extremamente sensíveis tanto a escolhas durante sua execução quanto variações nos dados. Se, por exemplo, removermos um subconjunto aleatório de observações do nosso dataset podemos obter aglomerações consideravelmente diferentes.

Portanto, agrupamentos devem ser usados como indicadores de tendências e estudos complementares aos métodos de pesquisa e experimentação mais rigorosos.

Uma análise de grupos pode demonstrar para um pesquisador uma possível relação entre certas observações e suas variáveis e cabe a ele investigar e comprovar essa relação mais a fundo.

Prática:

Carregamos o dataset USArrests e a biblioteca da qual ele depende.

```
library(Datasets)
```

```
crimes <- USArrests
```

O conjunto contempla todos os estados americanos e as variáveis 1, 2 e 4 se referem aos respectivos a cada 100 mil habitantes e a variável 3 é a porcentagem da população do estado que reside em centros urbanos.

	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7
Connecticut	3.3	110	77	11.1
Delaware	5.9	238	72	15.8
Florida	15.4	335	80	31.9
Georgia	17.4	211	60	25.8
Hawaii	5.3	46	83	20.2
Idaho	2.6	120	54	14.2
Illinois	10.4	249	83	24.0
Indiana	7.2	113	65	21.0

Prática:

Normalizamos o conjunto com a função `scale()` que retorna o dataset normalizado:

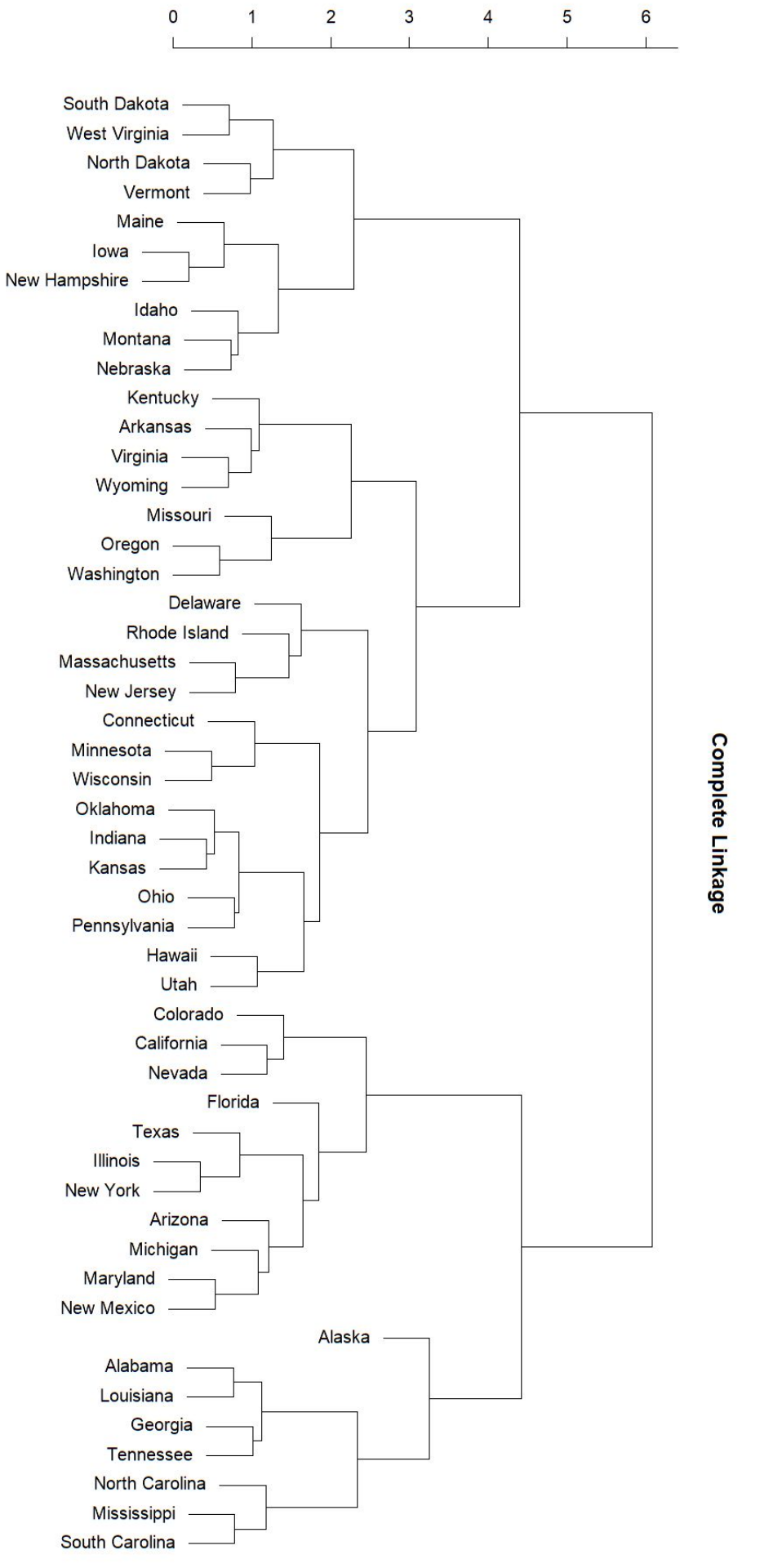
```
crimes_scale <- scale(crimes)
```

Em seguida geramos a matriz de distâncias com dist() que utiliza a distância euclidiana por padrão.

```
crimes_dist <- dist(crimes_scale)
```

Agora já podemos plotar o dendrograma para este conjunto a partir da seguinte chamada de função:

```
plot(hclust(crimes_dist), xlab = "", sub = "", ylab = "",
    labels = row.names(crimes), main = "Complete Linkage")
```

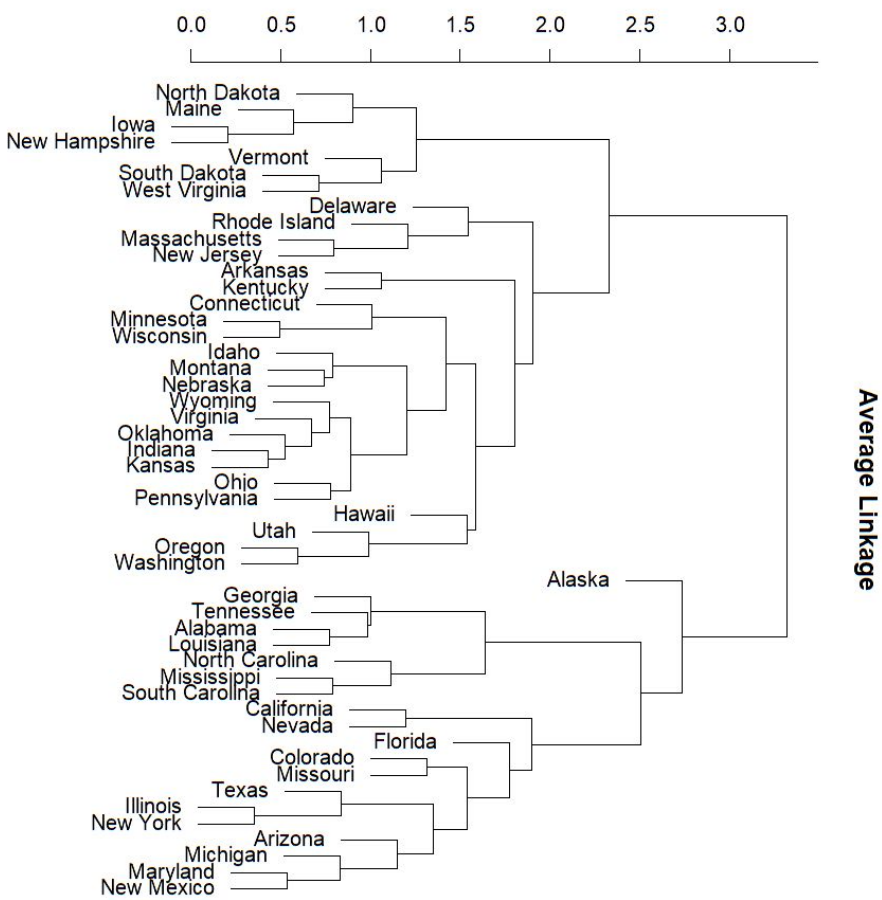

0
1
2
3
4
5
6
South Dakota
West Virginia
North Dakota
Vermont
Maine
Iowa
New Hampshire
Idaho
Montana
Nebraska
Kentucky
Arkansas
Virginia
Wyoming
Missouri
Oregon
Washington
Delaware
Rhode Island
Massachusetts
New Jersey
Connecticut
Minnesota
Wisconsin
Oklahoma
Indiana
Kansas
Ohio
Pennsylvania
Hawaii
Utah
Colorado
California
Nevada
Florida
Texas
Illinois
New York
Arizona
Michigan
Maryland
New Mexico
Alaska
Alabama
Louisiana
Georgia
Tennessee
North Carolina
Mississippi
South Carolina
Complete Linkage

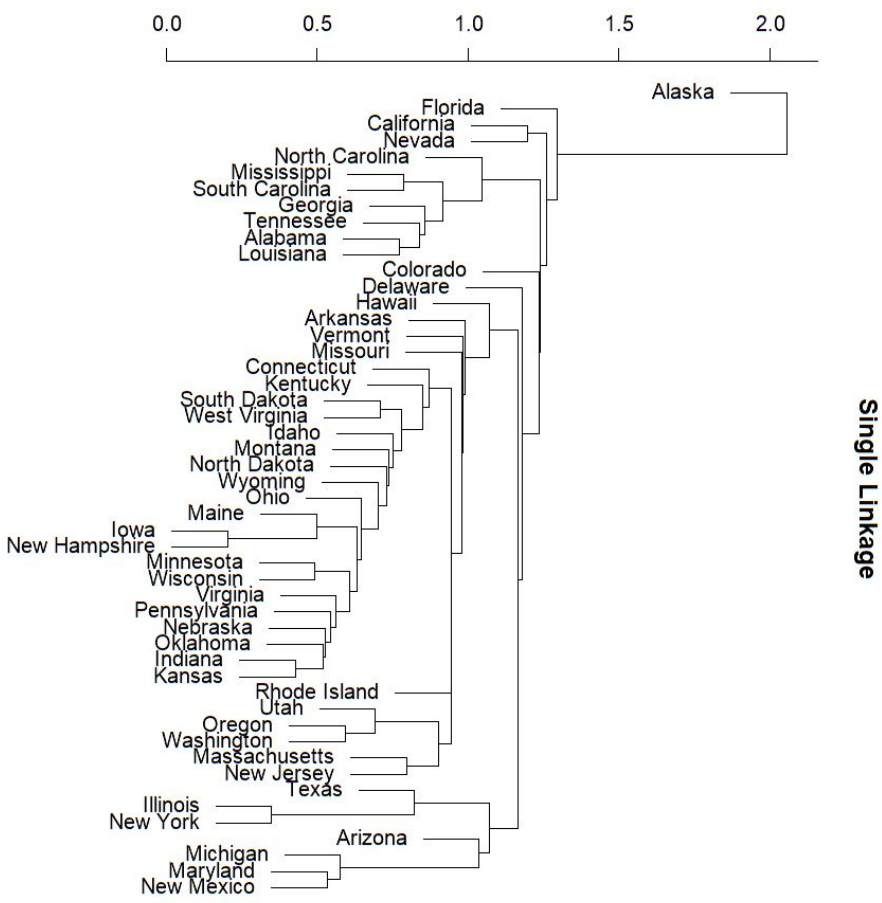
Prática:

A função `hclust()` foi responsável por gerar um “objeto dendrograma” (simplificação do funcionamento real do objeto) e este foi plotado pela função `plot()`.

Podemos gerar agrupamentos hierárquicos das outras formas de linkagem alterando o campo implicito method.

```
plot(hclust(crimes_dist, method = "average"), xlab = "",
    sub = "", ylab = "", labels = row.names(crimes),
    main = "Average Linkage")
```


0.0
0.5
1.0
1.5
2.0
2.5
3.0
North Dakota
Maine
Iowa
New Hampshire
Vermont
South Dakota
West Virginia
Delaware
Rhode Island
Massachusetts
New Jersey
Arkansas
Kentucky
Connecticut
Minnesota
Wisconsin
Idaho
Montana
Nebraska
Wyoming
Virginia
Oklahoma
Indiana
Kansas
Ohio
Pennsylvania
Hawaii
Utah
Oregon
Washington
Alaska
Georgia
Tennessee
Alabama
Louisiana
North Carolina
Mississippi
South Carolina
California
Nevada
Florida
Colorado
Missouri
Texas
Illinois
New York
Arizona
Michigan
Maryland
New Mexico
Average Linkage


0.0
0.5
1.0
1.5
2.0
Alaska
Florida
California
Nevada
North Carolina
Mississippi
South Carolina
Georgia
Tennessee
Alabama
Louisiana
Colorado
Delaware
Hawaii
Arkansas
Vermont
Missouri
Connecticut
Kentucky
South Dakota
West Virginia
Idaho
Montana
North Dakota
Wyoming
Ohio
Maine
Iowa
New Hampshire
Minnesota
Wisconsin
Virginia
Pennsylvania
Nebraska
Oklahoma
Indiana
Kansas
Rhode Island
Utah
Oregon
Washington
Massachusetts
New Jersey
Texas
Illinois
New York
Arizona
Michigan
Maryland
New Mexico
Single Linkage

Prática:

Com um “objeto dendrograma” podemos também chamar funções auxiliares para analisar seus componentes.

```
crimes_dend <- hclust(crimes_dist)
```

```
grupos <- cutree(crimes_dend, 5)
```

A função `cutree()` divide o dendrograma em *x* grupos e retorna um vetor do tamanho da quantidade de observações onde cada posição *i* armazena a qual grupo a *i-ésima* observação pertence.

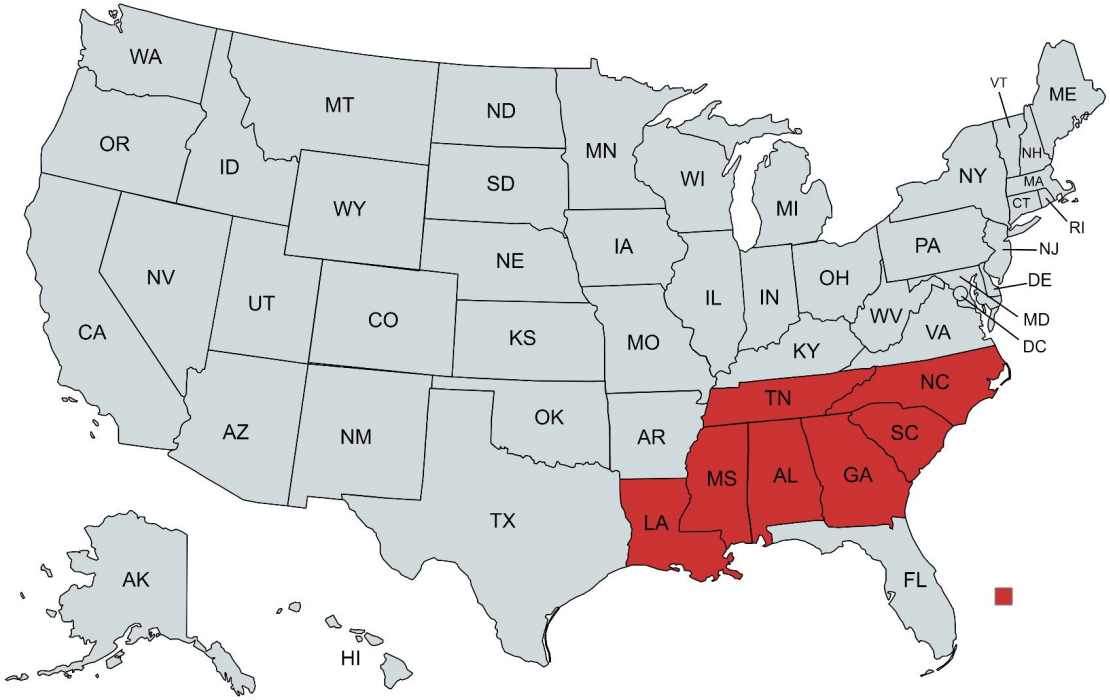
A partir dos grupos, podemos extrair os nomes dos estados que pertencem a um certo grupo:

```
row.names(crimes)[grupos==1]
```

Prática:

Utilizando um software online podemos pintar no mapa os estados presentes em um determinado grupo.

Neste caso o grupo escolhido foi o 1.



Prática:

Também podemos usar dos grupos para formar subsets do conjunto de dados original e a partir destes analisar o comportamento das variáveis.

```
grupo1 <- crimes[grupos==1,]

grupo3 <- crimes[grupos==3,]

summary(grupo1)

summary(grupo3)
```

E obtemos as seguintes informações, cuja interpretação é deixada como um exercício para o leitor:

Prática:

```
> summary(grupo1)
     Murder          Assault         UrbanPop          Rape      
 Min.   :13.00   Min.   :188.0   Min.   :44.00   Min.   :16.10  
 1st Qu.:13.20   1st Qu.:223.5   1st Qu.:46.50   1st Qu.:19.15  
 Median :14.40   Median :249.0   Median :58.00   Median :22.20  
 Mean   :14.67   Mean   :251.3   Mean   :54.29   Mean   :21.69  
 3rd Qu.:15.75   3rd Qu.:269.0   3rd Qu.:59.50   3rd Qu.:24.15  
 Max.   :17.40   Max.   :337.0   Max.   :66.00   Max.   :26.90  
> summary(grupo3)
     Murder          Assault         UrbanPop          Rape      
 Min.   : 7.90   Min.   :201.0   Min.   :67.00   Min.   :24.00  
 1st Qu.: 9.70   1st Qu.:250.5   1st Qu.:76.00   1st Qu.:26.95  
 Median :11.30   Median :255.0   Median :80.00   Median :31.90  
 Mean   :11.05   Mean   :264.1   Mean   :79.09   Mean   :32.62  
 3rd Qu.:12.15   3rd Qu.:289.5   3rd Qu.:82.00   3rd Qu.:36.90  
 Max.   :15.40   Max.   :335.0   Max.   :91.00   Max.   :46.00  
```

Prática:

Até então abordamos apenas agrupamento hierárquico, porém o K-médias também é computável:

```
set.seed(sample(1:1000,1))
```

```
crimes_km <- kmeans(crimes_scale, 5)
```

Devemos utilizar `set.seed()` pois o método depende da aleatoriedade. A função `kmeans()` recebe os dados e a quantidade de grupos e retorna um “objeto k-médias”.

A mesma lógica do método anterior se aplica, podendo armazenar os grupos em um vetor:

```
grupos_km <- crimes_km$cluster
```

Prática:

Aplicando as mesmas funções de antes (slide 35), podemos obter as seguintes informações:

```
> summary(grupo_km1)
     Murder          Assault         UrbanPop          Rape      
 Min.   : 8.8   Min.   :188.0   Min.   :50.0   Min.   :19.50  
 1st Qu.:13.2   1st Qu.:190.0   1st Qu.:58.0   1st Qu.:21.20  
 Median :13.2   Median :211.0   Median :59.0   Median :22.20  
 Mean   :13.6   Mean   :214.8   Mean   :58.6   Mean   :23.12  
 3rd Qu.:15.4   3rd Qu.:236.0   3rd Qu.:60.0   3rd Qu.:25.80  
 Max.   :17.4   Max.   :249.0   Max.   :66.0   Max.   :26.90  
> summary(grupo_km3)
     Murder       Assault       UrbanPop        Rape      
 Min.   :10   Min.   :263   Min.   :48   Min.   :44.5  
 1st Qu.:10   1st Qu.:263   1st Qu.:48   1st Qu.:44.5  
 Median :10   Median :263   Median :48   Median :44.5  
 Mean   :10   Mean   :263   Mean   :48   Mean   :44.5  
 3rd Qu.:10   3rd Qu.:263   3rd Qu.:48   3rd Qu.:44.5  
 Max.   :10   Max.   :263   Max.   :48   Max.   :44.5
```

Prática:

E, por fim, podemos comparar os agrupamentos dos dois métodos:

```
table(grupos,grupos_km)
```

```
> table(grupos_km, grupos)
          grupos
grupos_km  1  2  3  4  5
        1  0  0  0  9  2
        2  0  0  0 10  0
        3  0  0  0  2  8
        4  7  0  0  0  0
        5  0  1 11  0  0
```

Embora interessante, essa comparação não é muito útil pois podemos acabar tendo grupos parecidos porém associados a números diferentes.

A comparação correta implica em analisar um a um os grupos formados em um dos métodos e compará-lo com todos do outro método a partir dos nomes de suas observações.

Conclusão:

Métodos de agrupamento são ferramentas que permitem uma boa visualização do comportamento de observações e suas variáveis assim como a semelhança entre elas.

Devemos tomar uma quantidade significativa de decisões quando escolhemos e implementamos um método, tornando esse processo extremamente subjetivo e dependente dos conhecimentos interdisciplinares do pesquisador.

Estes métodos não servem como foco central de uma análise e devem ser usados como ponto de partida para guiar um pesquisador em uma determinada direção ou como ferramenta para reforçar uma tese já feita com embasamento científico.

FIM.