Análise de agrupamento

Patrícia de Siqueira Ramos

UNIFAL-MG, *campus* Varginha

9 de Junho de 2016

Técnicas hierárquicas e seleção de variáveis

- Os métodos hierárquicos aglomerativos de análise de agrupamento também podem ser úteis na seleção das variáveis mais importantes para caracterizar uma situação
- Assim, se o interesse do pesquisador for agrupar as variáveis que forem mais similares entre si e separar aquelas que tenham informações diferenciadas, deve-se escolher uma matriz inicial que represente o relacionamento dessas variáveis

Técnicas hierárquicas e seleção de variáveis

- Os métodos hierárquicos aglomerativos de análise de agrupamento também podem ser úteis na seleção das variáveis mais importantes para caracterizar uma situação
- Assim, se o interesse do pesquisador for agrupar as variáveis que forem mais similares entre si e separar aquelas que tenham informações diferenciadas, deve-se escolher uma matriz inicial que represente o relacionamento dessas variáveis
- Para variáveis quantitativas, a medida de relacionamento mais natural é o coeficiente de correlação
- A matriz de correlação, por si só, não é uma matriz de distâncias, mas transformações podem ser feitas. Uma delas é

$$\mathbf{D}_{p\times p} = \mathbf{1}_{p\times p} - abs(\mathbf{R}_{p\times p}).$$

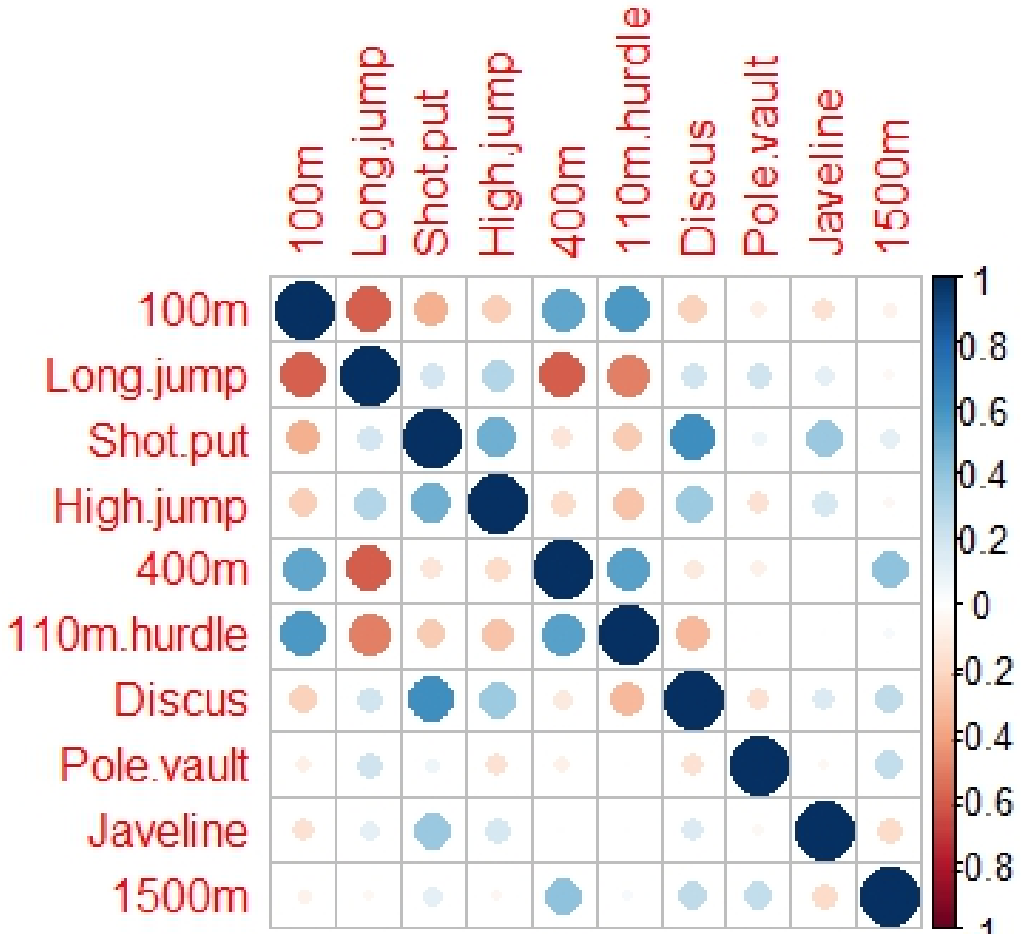
Técnicas hierárquicas e seleção de variáveis

- A matriz **D** caracteriza a proximidade entre as variáveis e serviria para implementar os métodos de agrupamento
- Assim, as variáveis em um mesmo grupo seriam altamente correlacionadas entre si e, as de grupos diferentes, seriam pouco correlacionadas entre si
- Para a seleção final de variáveis que seriam usadas, poderiam ser escolhidas, dentro de cada grupo formado, algumas variáveis

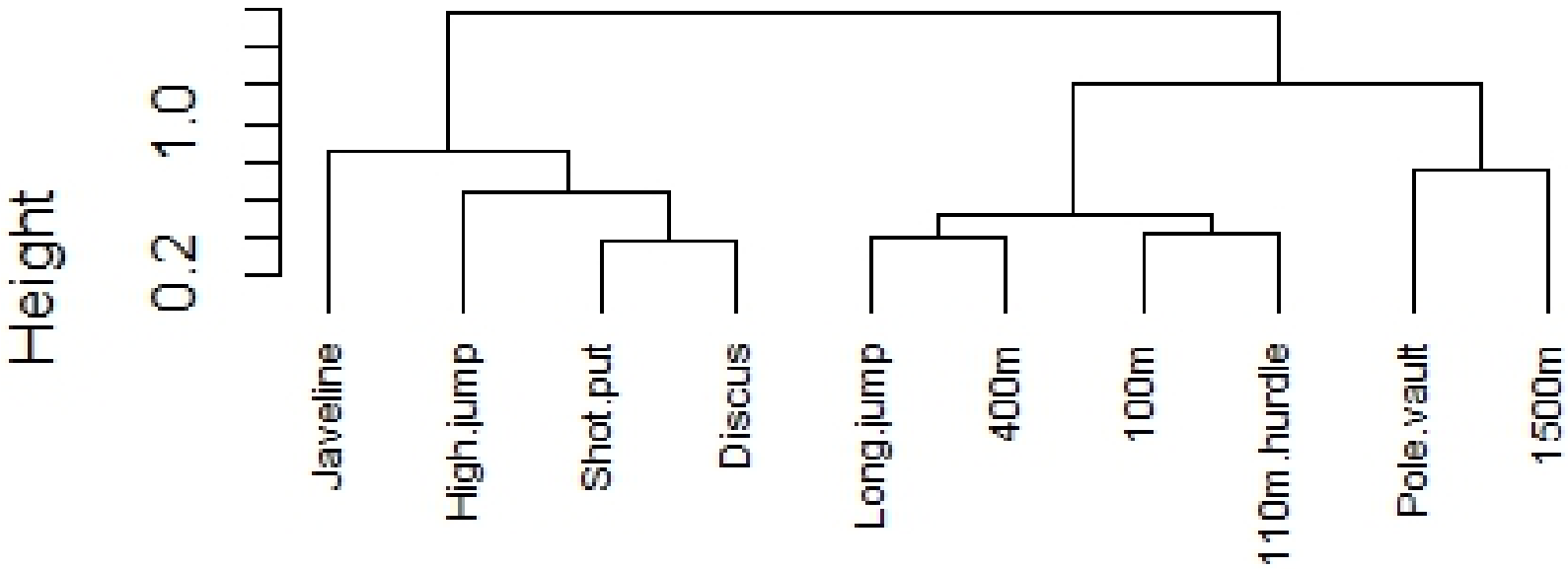
Ex.: Dados de decatlon. Matriz de correlações:

	$100m$	*Long.j*	*Shot.p*	*High.j*	$400m$	$110m.h$	*Discus*	*Pole.v*	*Jav*	$1500m$
$100m$	1,00	−0,60	−0,36	−0,25	0,52	0,58	−0,22	−0,08	−0,16	−0,06
Long.j	−0,60	1,00	0,18	0,29	−0,60	−0,51	0,19	0,20	0,12	−0,03
Shot.p	−0,36	0,18	1,00	0,49	−0,14	−0,25	0,62	0,06	0,37	0,12
High.j	−0,25	0,29	0,49	1,00	−0,19	−0,28	0,37	−0,16	0,17	−0,04
$400m$	0,52	−0,60	−0,14	−0,19	1,00	0,55	−0,12	−0,08	0,00	0,41
$110m.h$	0,58	−0,51	−0,25	−0,28	0,55	1,00	−0,33	0,00	0,01	0,04
Discus	−0,22	0,19	0,62	0,37	−0,12	−0,33	1,00	−0,15	0,16	0,26
Pole.v	−0,08	0,20	0,06	−0,16	−0,08	0,00	−0,15	1,00	−0,03	0,25
Jav	−0,16	0,12	0,37	0,17	0,00	0,01	0,16	−0,03	1,00	−0,18
$1500m$	−0,06	−0,03	0,12	−0,04	0,41	0,04	0,26	0,25	−0,18	1,00

Outra forma de representar as correlações (pacote `corrplot` do R)



Agrupamento das variáveis dos dados de decatlon



b) Métodos de agrupamento não hierárquicos

- Objetivo: particionar n observações em k grupos de modo que haja homogeneidade dentro dos grupos e heterogeneidade entre os grupos formados
- O número k de grupos deve ser escolhido *a priori*

b) Métodos de agrupamento não hierárquicos

- Objetivo: particionar n observações em k grupos de modo que haja homogeneidade dentro dos grupos e heterogeneidade entre os grupos formados
- O número k de grupos deve ser escolhido *a priori*
- Para encontrar a "melhor" partição, algum critério de qualidade da partição deve ser empregado
- Impossível obter e avaliar todas as partições possíveis de ordem k e, por isso, avaliam-se apenas algumas para obter a "quase ótima"

b) Métodos de agrupamento não hierárquicos

- Nos métodos hierárquicos, o processo é aplicado à matriz de distâncias e, nos métodos não hierárquicos, o processo é aplicado à matriz de dados
- Não é possível construir dendrogramas pois observações podem entrar e sair de grupos em qualquer passo

b) Métodos de agrupamento não hierárquicos

- Nos métodos hierárquicos, o processo é aplicado à matriz de distâncias e, nos métodos não hierárquicos, o processo é aplicado à matriz de dados
- Não é possível construir dendrogramas pois observações podem entrar e sair de grupos em qualquer passo
- Métodos comuns:
  - k-médias
  - Redes neurais aplicadas à análise de agrupamento
  - *Fuzzy* c-médias

i) Método das k-médias

- É um dos métodos mais conhecidos e utilizados em problemas práticos
- Basicamente, cada observação é alocada àquele grupo cujo centróide (vetor de médias amostral) é o mais próximo do vetor de valores observados para o respectivo elemento

i) Método das k-médias

- É um dos métodos mais conhecidos e utilizados em problemas práticos
- Basicamente, cada observação é alocada àquele grupo cujo centróide (vetor de médias amostral) é o mais próximo do vetor de valores observados para o respectivo elemento
- A implementação mais utilizada é a que procura a partição das n observações em k grupos que minimize a soma de quadrados dentro dos grupos (SQDG):

$$\text{SQDG} = \sum_{j=1}^{p} \sum_{l=1}^{k} \sum_{i \in G_l} (X_{ij} - \bar{X}_j^{(l)})^2,$$

em que $\bar{X}_j^{(l)} = \frac{1}{n_i} \sum_{i \in G_l} X_{ij}$ é a média das observações do grupo G_l na variável j.

Implementação do método das k-médias

Uma das implementações do método das k-médias tem os seguintes passos:

1. Alocar arbitrariamente os n objetos aos k grupos e calcular os seus centróides. Pode-se gerar os centróides de cada grupo por um processo aleatório qualquer

Implementação do método das k-médias

Uma das implementações do método das k-médias tem os seguintes passos:

1. Alocar arbitrariamente os n objetos aos k grupos e calcular os seus centróides. Pode-se gerar os centróides de cada grupo por um processo aleatório qualquer
2. Alocar cada um dos n objetos aos grupos que apresentem a menor distância (geralmente euclidiana) com o respectivo objeto (minimização da SQDG)

Implementação do método das k-médias

Uma das implementações do método das k-médias tem os seguintes passos:

1. Alocar arbitrariamente os n objetos aos k grupos e calcular os seus centróides. Pode-se gerar os centróides de cada grupo por um processo aleatório qualquer
2. Alocar cada um dos n objetos aos grupos que apresentem a menor distância (geralmente euclidiana) com o respectivo objeto (minimização da SQDG)
3. Recalcular os centróides de cada grupo

Implementação do método das k-médias

Uma das implementações do método das k-médias tem os seguintes passos:

1. Alocar arbitrariamente os n objetos aos k grupos e calcular os seus centróides. Pode-se gerar os centróides de cada grupo por um processo aleatório qualquer
2. Alocar cada um dos n objetos aos grupos que apresentem a menor distância (geralmente euclidiana) com o respectivo objeto (minimização da SQDG)
3. Recalcular os centróides de cada grupo
4. Realocar o primeiro objeto de seu grupo para um outro grupo em que a distância for mínima e menor do que a distância desse objeto para seu próprio grupo de origem

Implementação do método das k-médias

Uma das implementações do método das k-médias tem os seguintes passos:

1. Alocar arbitrariamente os n objetos aos k grupos e calcular os seus centróides. Pode-se gerar os centróides de cada grupo por um processo aleatório qualquer
2. Alocar cada um dos n objetos aos grupos que apresentem a menor distância (geralmente euclidiana) com o respectivo objeto (minimização da SQDG)
3. Recalcular os centróides de cada grupo
4. Realocar o primeiro objeto de seu grupo para um outro grupo em que a distância for mínima e menor do que a distância desse objeto para seu próprio grupo de origem
5. Repetir os passos 3 e 4 até que não ocorram mais mudanças de objetos de um grupo para outro (é realizada apenas uma transferência por iteração)

Opções para inicialização do procedimento das k-médias

- Alocar arbitrariamente os n objetos aos k grupos
- Utilizar sementes aleatórias para representar o centróide dos k grupos
- Amostrar k objetos entre os n originais para representar inicialmente as sementes ou centróides dos grupos

Opções para inicialização do procedimento das k-médias

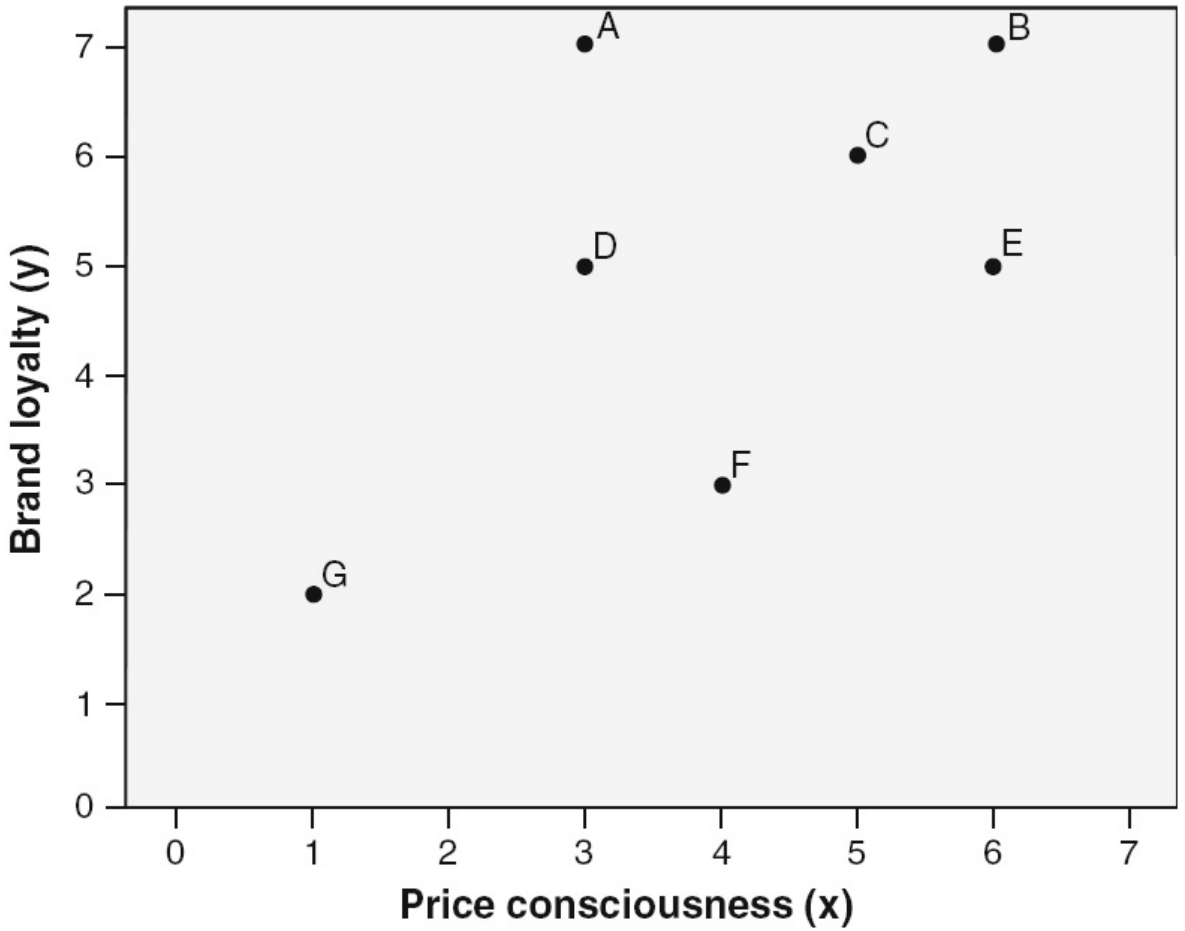
- Alocar arbitrariamente os n objetos aos k grupos
- Utilizar sementes aleatórias para representar o centróide dos k grupos
- Amostrar k objetos entre os n originais para representar inicialmente as sementes ou centróides dos grupos

Observações

- O método é sensível à escolha inicial dos grupos ou de seus centróides
- Aconselhável repetir o processo todo com escolhas diferentes das sementes iniciais
- Se as diferentes escolhas levarem a a grupos finais muito diferentes ou se a convergência for muito lenta, pode não haver grupos naturais no conjunto de dados

Exemplo de aplicação do método das k-médias

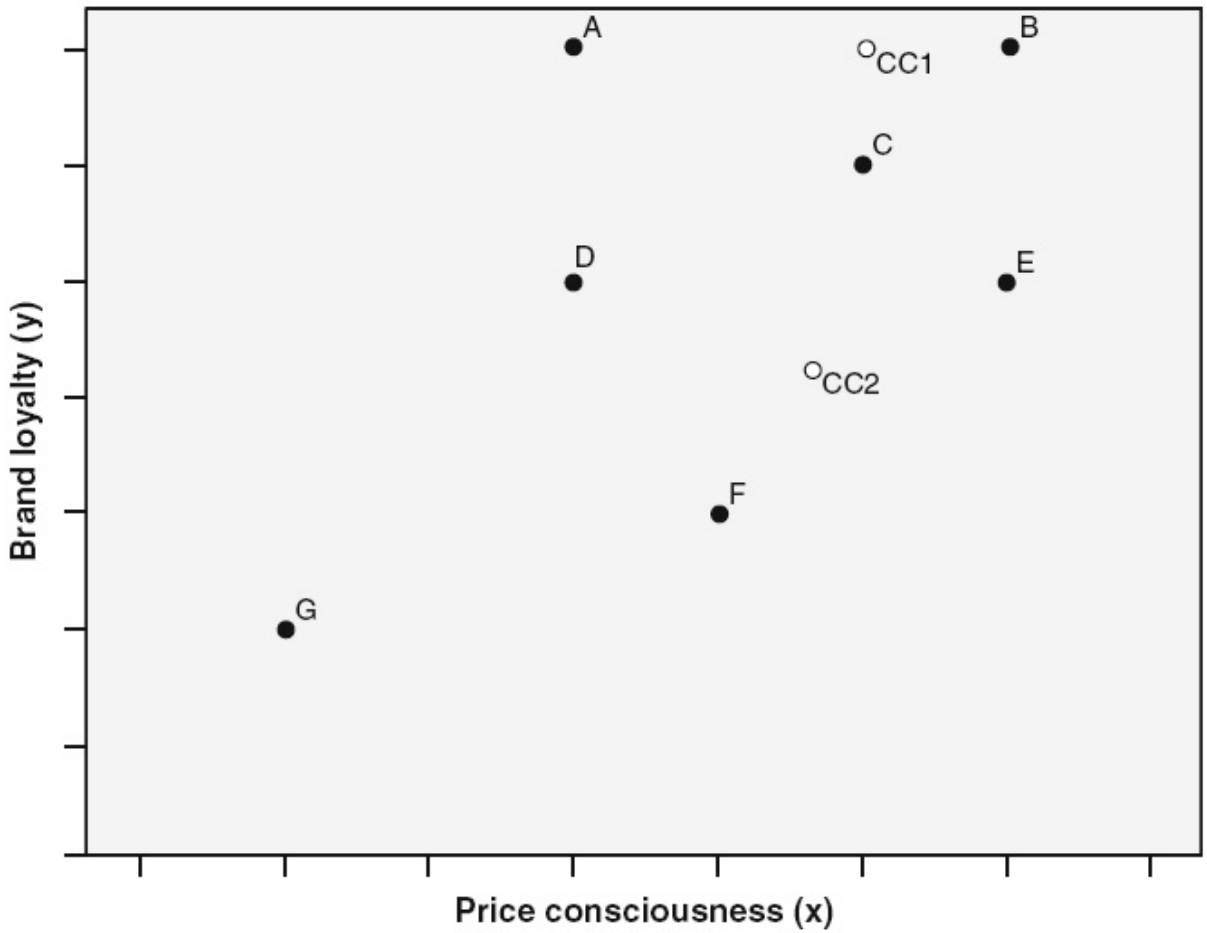
O mercado de uma marca será separado de acordo com duas variáveis: X - consciência sobre preço e Y - fidelidade à marca.



Exemplo de aplicação do método das k-médias - Passo 1

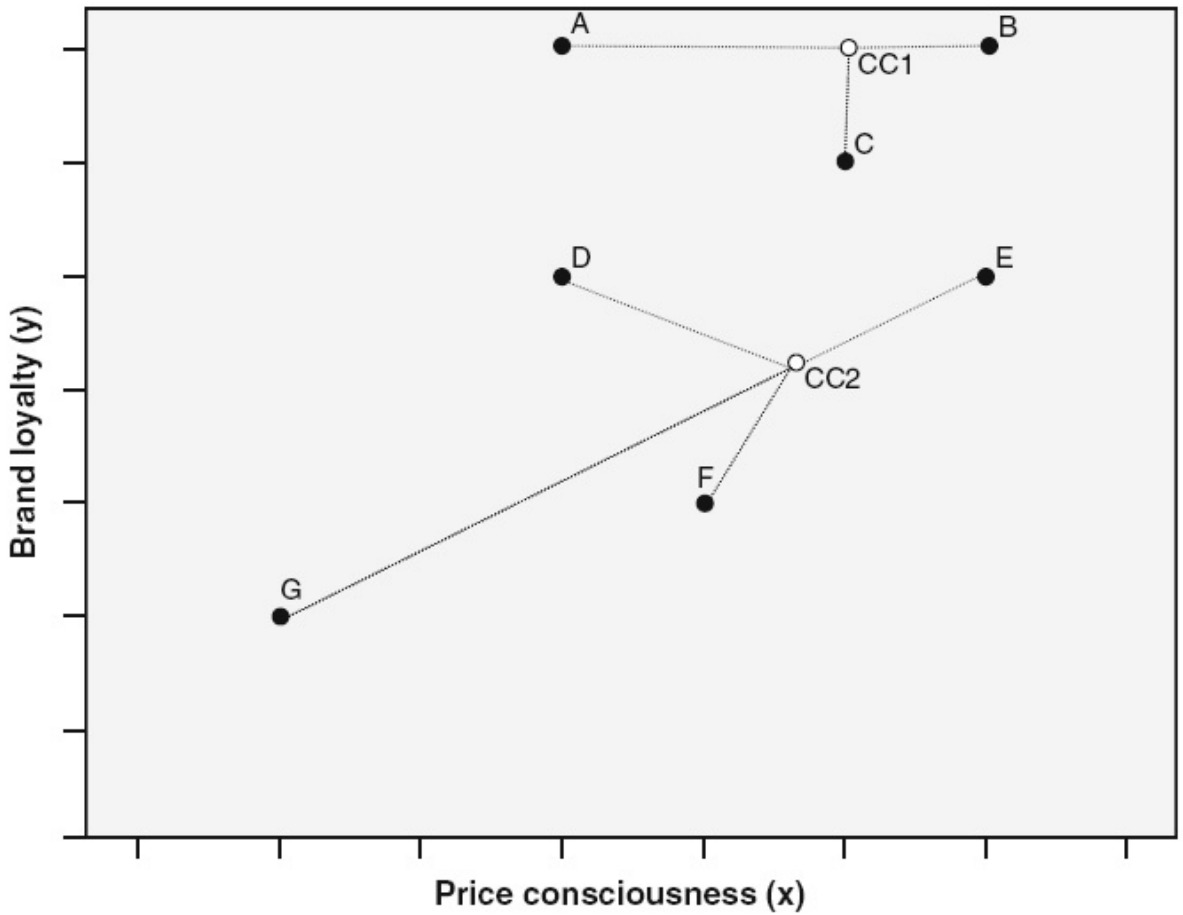
Valor de k definido como 2 (2 grupos de clientes).

O algoritmo seleciona aleatoriamente um centro para cada grupo - $CC1$ e $CC2$.



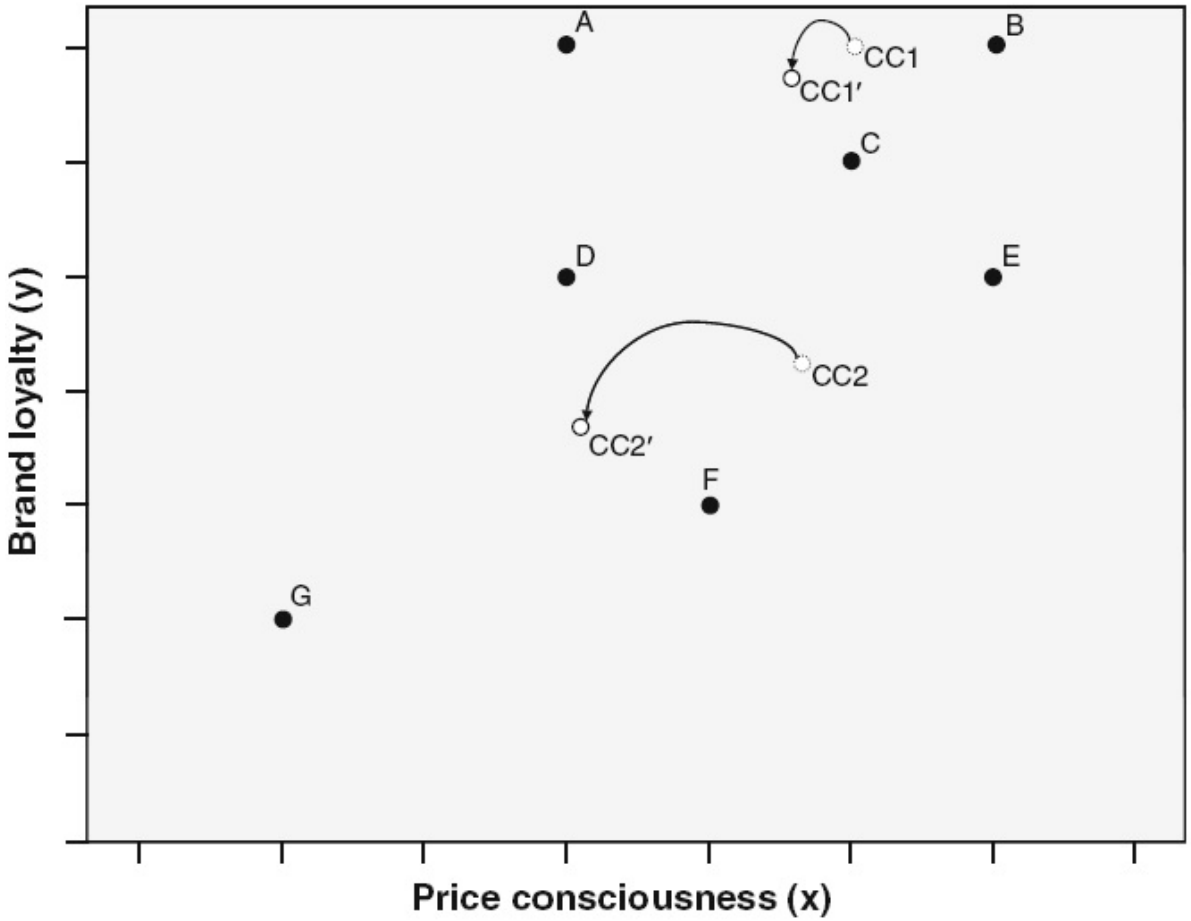
Exemplo de aplicação do método das *k*-médias - Passo 2

Distâncias euclidianas são calculadas dos centros para cada obs. Cada obs. ficará associada ao centro para o qual a distância é menor (A, B, C: 1/ D, E, F, G: 2).



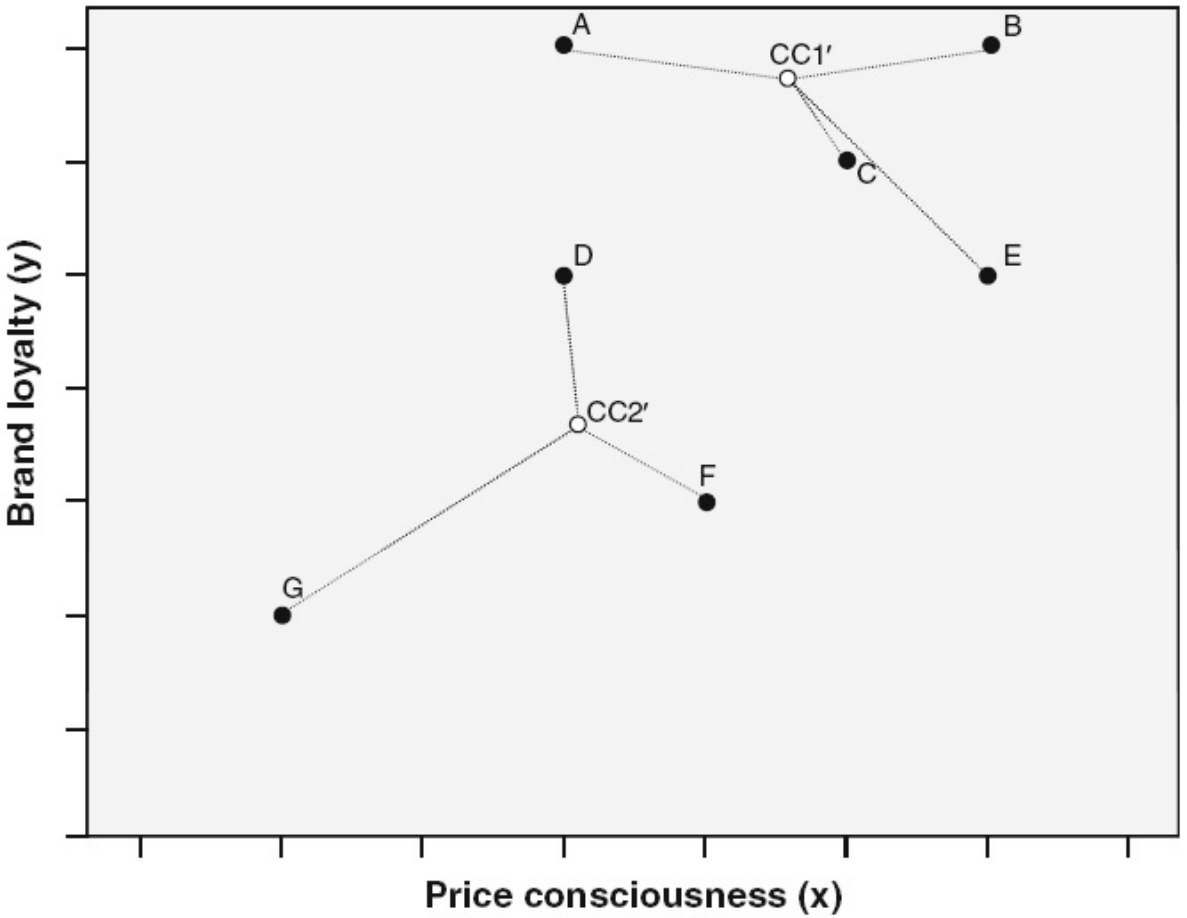
Exemplo de aplicação do método das k-médias - Passo 3

Baseando-se na partição realizada, cada centróide é calculado (valores médios das obs. em cada grupo). $CC1$ e $CC2$ mudam de lugar.



Exemplo de aplicação do método das *k*-médias - Passo 4

Distâncias das obs. para os novos centros são calculadas e as obs. são atribuídas aos grupos para os quais a distância para o centro é menor.



Resumo

- Os passos 3 e 4 são repetidos até atingir um número predefinido de iterações ou até a convergência ser atingida (não haver mais diferenças entre os grupos obtidos)
- Geralmente, o método das k-médias apresenta desempenho superior aos métodos hierárquicos e é menos afetado por *outliers*
- Pode ser aplicado a conjuntos de dados grandes ($n > 500$) e custa menos computacionalmente

Resumo

- Desvantagem: decidir o valor de k antes
- Solução 1: utilizar uma técnica hierárquica antes para definir o número k e, em seguida, aplicar o k-médias
- Solução 2: utilizar um gráfico da SQDG com a solução do método para cada número de grupos. Procura-se o "cotovelo" que poderá ajudar na decisão do valor de k
- Obs. (solução 2): à medida que k cresce, SQDG sempre decresce

Gráfico SQDG $\times$ k