Análise de agrupamento

Patrícia de Siqueira Ramos

UNIFAL-MG, *campus* Varginha

30 de Junho de 2018

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

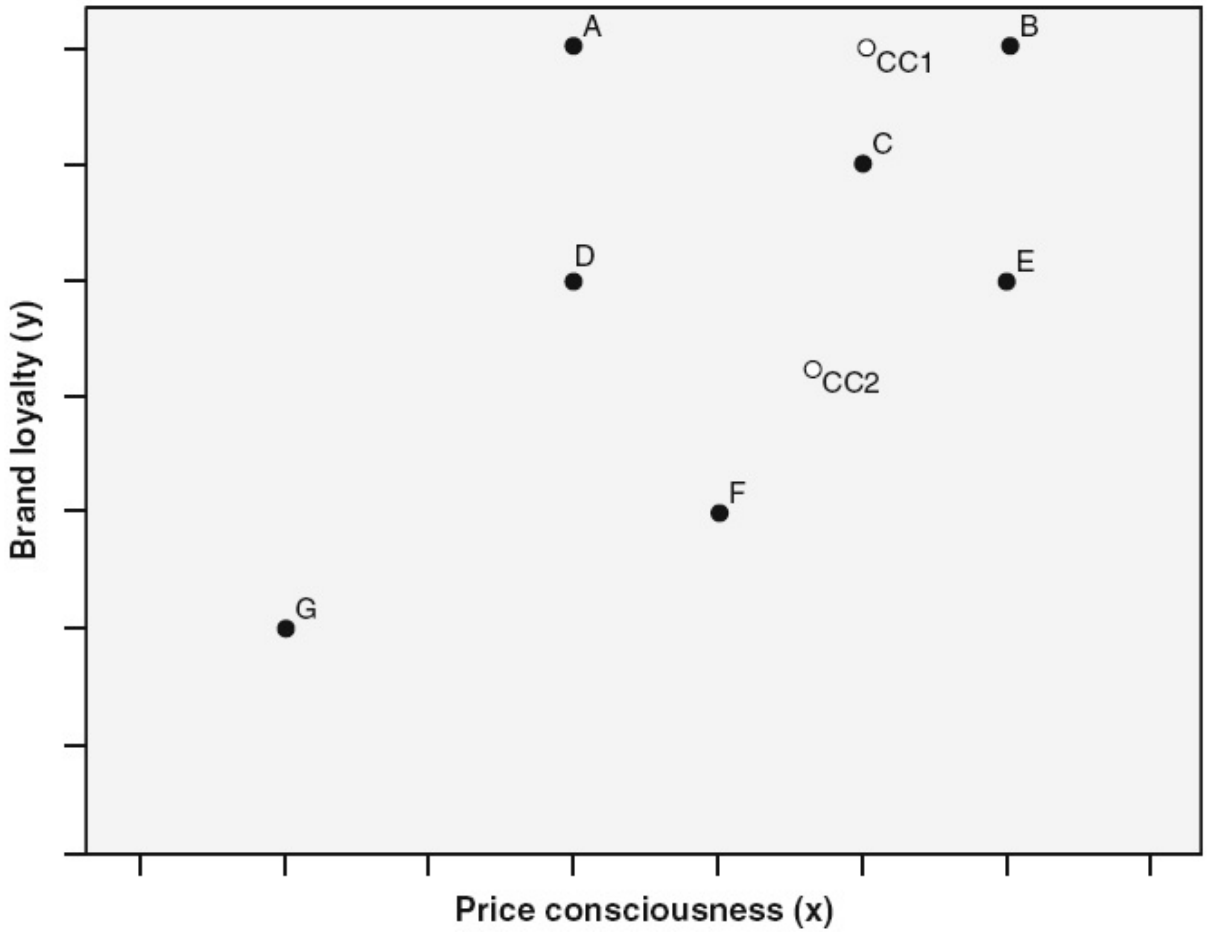
Exemplo de aplicação do método das k-médias

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

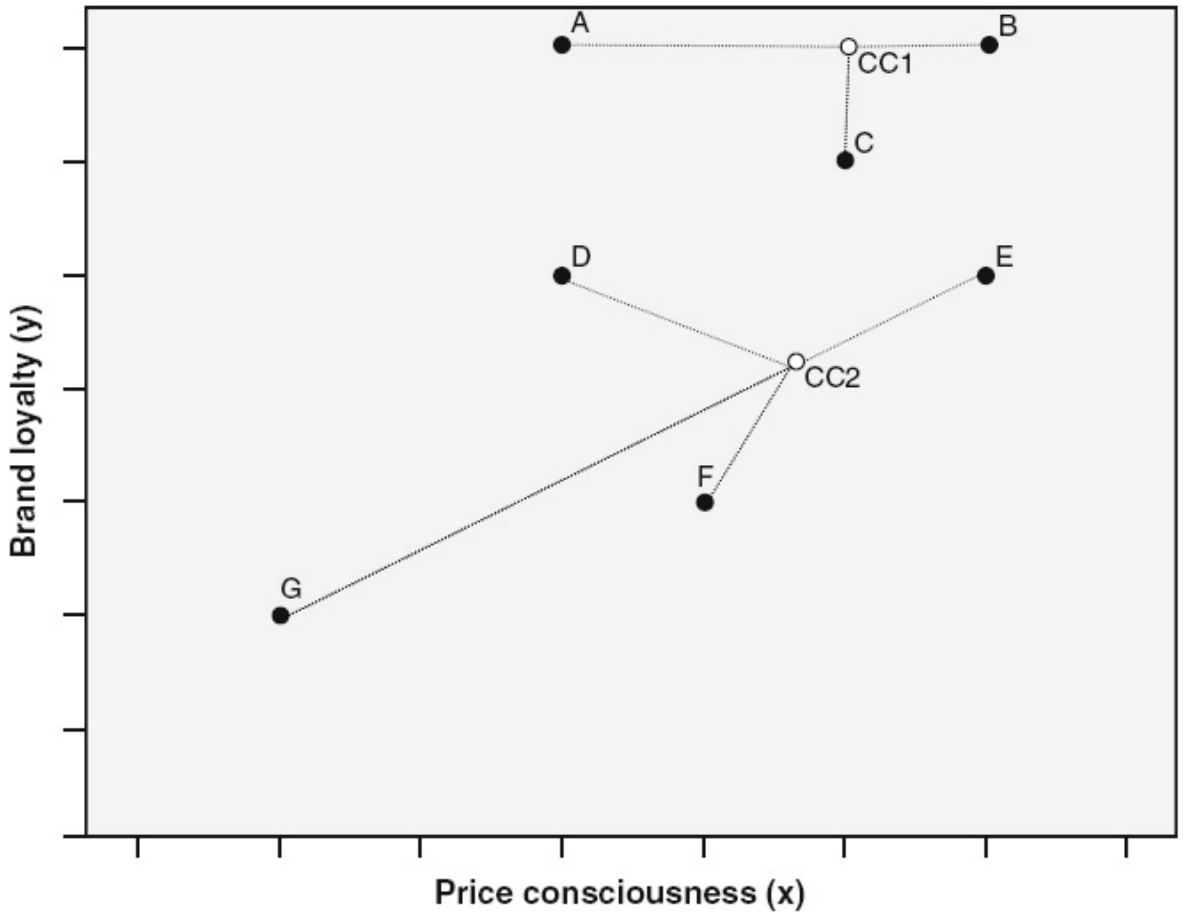
Valor de k definido como 2 (2 grupos de clientes).

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



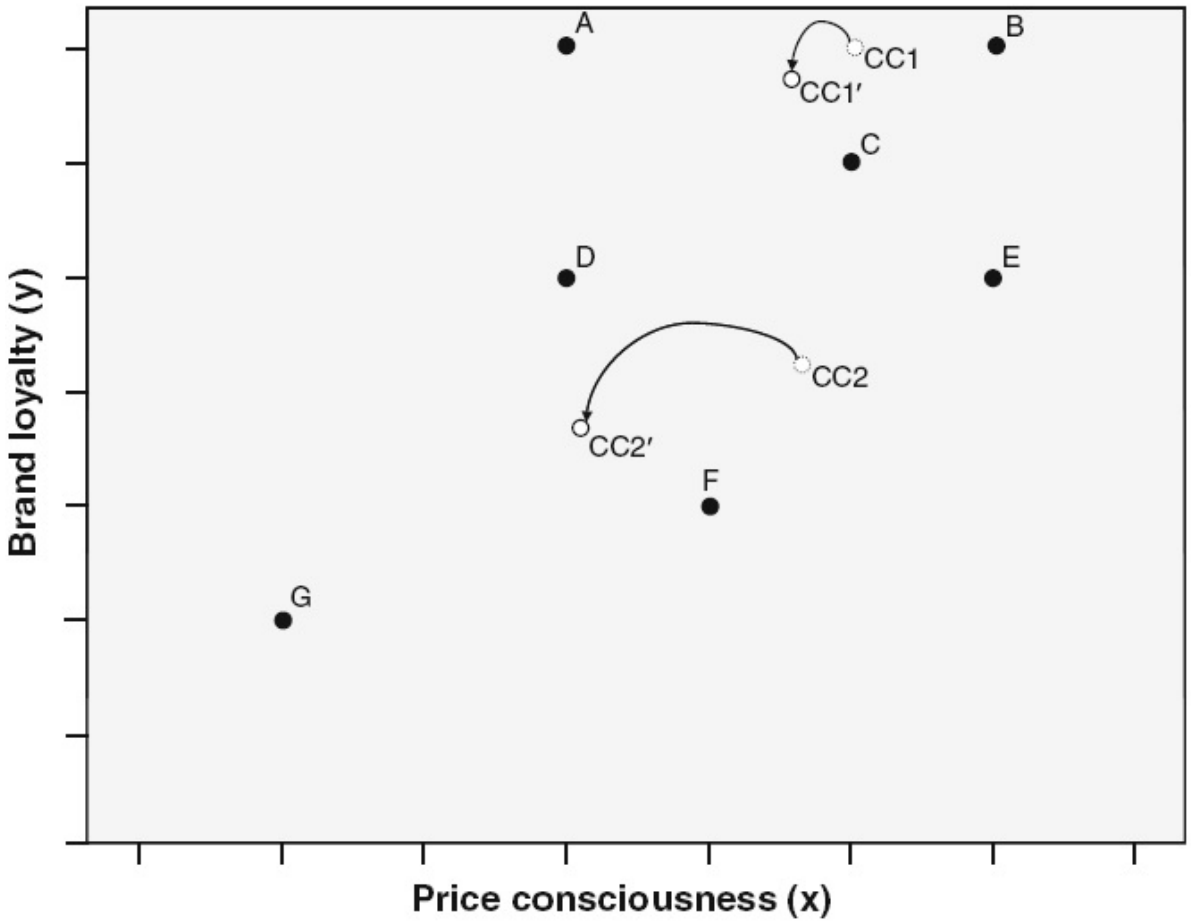
Exemplo de aplicação do método das *k*-médias - Passo 2

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).



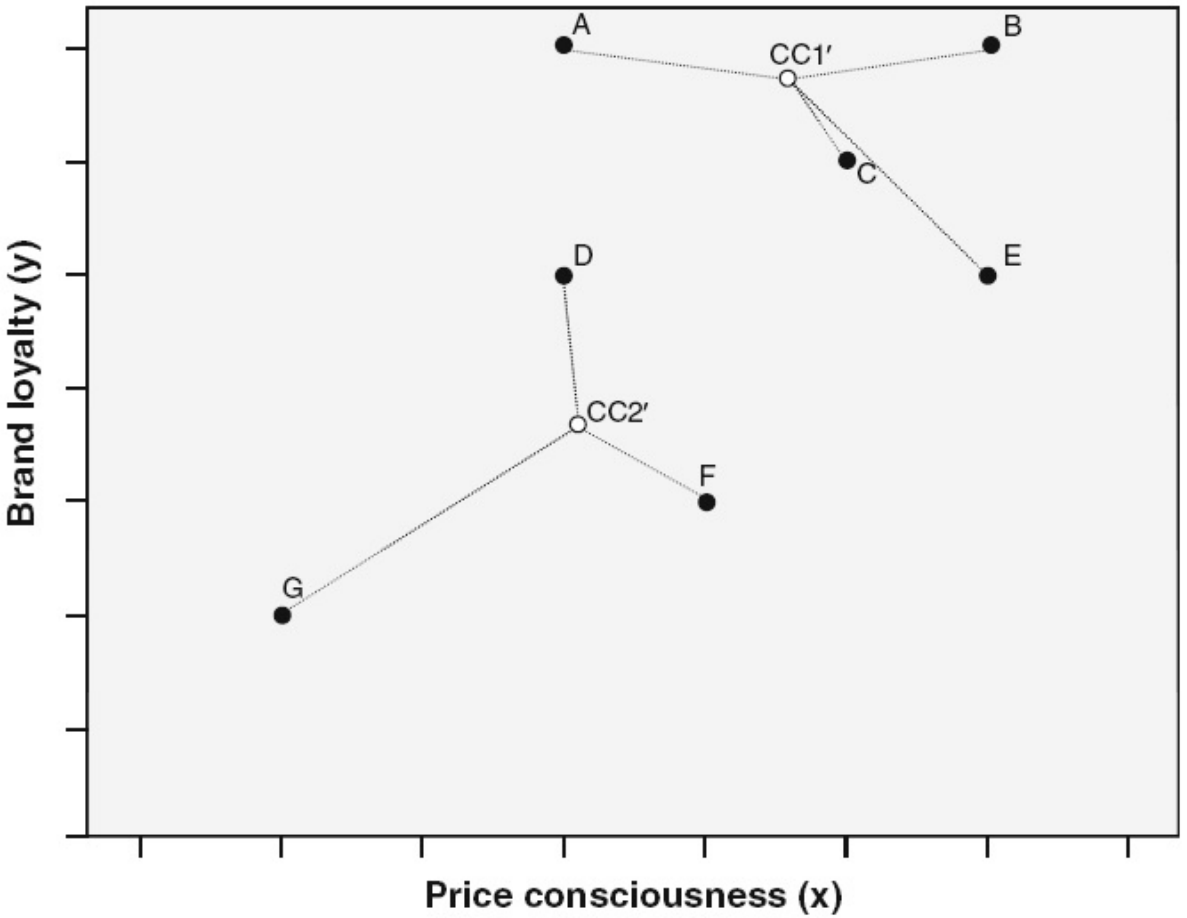
Exemplo de aplicação do método das k-médias - Passo 3

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.



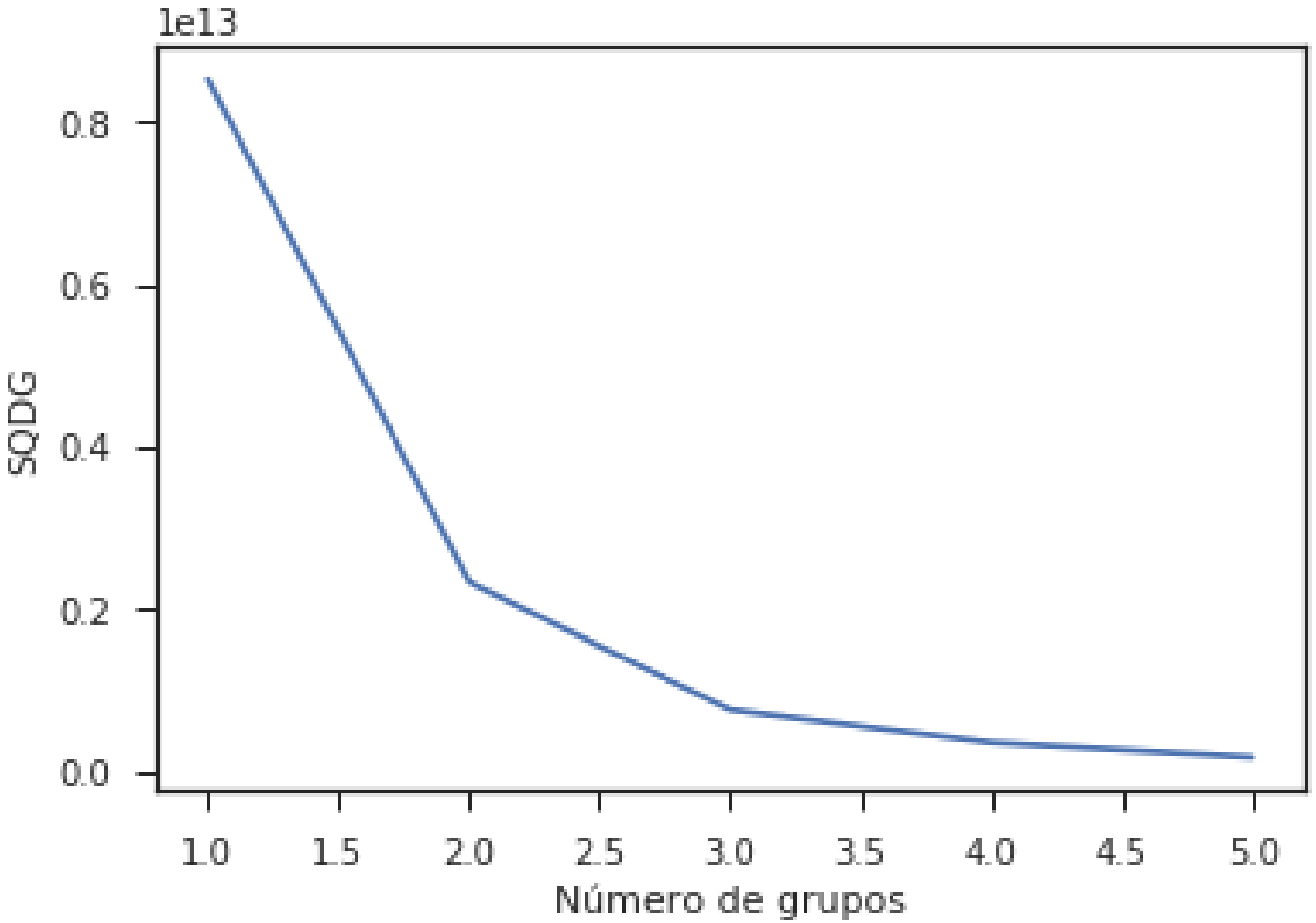
Resumo

- 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
- Solução 3: *silhouette plot* (a ser visto na aula prática)

Solução 2: Gráfico SQDG $\times$ k - ex. de municípios, p=4



Solução 2: Gráfico $k \times$ SQDG - ex. de municípios, p=4

- De acordo com o gráfico $k \times$ SQDG, percebe-se que a queda da SQDG se torna mais suave de 3 para 4 grupos
- Assim, a divisão em 3 grupos seria uma boa escolha
- Essa também foi a decisão baseando-se na solução 1, ou seja, usando uma técnica hierárquica (corte do dendrograma), também optamos por $k = 3$