Projeto de disciplina - Análise de clusters
Feito por: Rodrigo Bragança de Oliveira

Primeiramente apresento as perguntas, e as respondo após um “Resp:” em formato BOLD (NEGRITO), logo após as perguntas.

1. Para as questões a seguir você irá utilizar o ambiente RStudio (cloud ou local). Tire um printscreen da tela, para mostrar o ambiente operando. Nesse print, deixe claro a versão de R utilizada e do pacote 'factoextra'. **Resp: Feito! Print dentro da pasta**

2. Da definição da Wikipedia: "[...] quantização de cores ou quantização de imagem colorida é quantização aplicada em espaço de cores. Esse é um processo que reduz o número de cores distintas usadas em uma imagem, normalmente com a intenção de que a nova imagem possivelmente deva ficar visualmente similar à imagem original.".

   Vimos que a clusterização das tonalidades dos pixels, é uma maneira de reduzir o número de cores usadas na representação da imagem através do uso dos centróides. A imagem abaixo (artigo Finding Lena, the Patron Saint of JPEGs) é muito utilizada em estudos com processamento de imagens. Use o algoritmo CLARA nesta imagem e a represente com 3, 5 e 10 cores distintas. **Resp: Feita a análise no R, porém segue abaixo as imagens com 3, 5 e 10 cores distintas apenas.**

   3 cores:

5 cores:

10 cores:

Escreva suas conclusões referentes a qualidade de representação da imagem (a informação é da imagem é perdida? Melhora com mais cores?)

Resp: Com apenas 3 cores, a imagem de fato fica um pouco ruim de se visualizar. Com 5 a 10 cores distintas, já percebemos que a imagem se aproxima bastante uma da outra e também em relação à original. Contudo, se for bem no detalhe, é

possível perceber que a imagem original é mais "*smooth*", devido à grande variedade de cores. Especialmente em relação às sombras.

3. Escolha uma base de dados para realizar esse projeto. Essa base de dados será utilizada durante toda sua análise. Essa base necessita ter 4 (ou mais) variáveis de interesse, onde todas são numéricas (confira com o professor a possibilidade de utilização de dados categóricos). Caso você tenha dificuldade para escolher uma base, o professor da disciplina irá designar para você. Explique qual o motivo para a escolha dessa base e aponte os resultados esperados através da análise.
   Resp: Segue o link da base utilizada, porém também incluí na pasta do projeto.
   https://archive.ics.uci.edu/dataset/476/buddymove+data+set
   Motivo da escolha é que a base de dados foi feita com o foco no estudo de clusterização, especialmente por se tratar de dados numéricos, por exemplo. E também por apresentar os requisitos solicitados na pergunta.

4. Para essa questão utilize o pacote factoextra e a base escolhida na questão anterior:
   1. O algoritmo de K-Médias usa a distância euclidiana para determinar a distância de um dado ao centróide que está sendo ajustado. Por que a distância euclidiana não é uma boa medida para dados com grande dimensionalidade? Indique uma distância mais apropriada.
      Resp: Não devemos utilizar a distância euclidiana: quando os dados são categóricos, quando há muitos outliers em nossa base de dados, quando há variáveis com escalas muito diferentes (por isso é importante normalizar os dados) e por último quando temos dados com uma alta dimensionalidade.

      Sobre a alta dimensionalidade, destacamos a "maldição da dimensionalidade", que pode ser explicada por: quanto maior é a nossa dimensionalidade, utilizando a distância euclidiana, a tendência é a de perder a sua capacidade discriminativa. Ou seja, nestes casos os pontos de dados tendem a ficar muito distantes uns dos outros, tornando as distâncias menos significativas entre eles. Com isso, é preferível utilizar outra medida de distância/similaridade (ex. Similaridade de Cosseno, que é recomendada para dados esparsos, ou seja, com alta dimensionalidade).

   2. A normalização dos dados é uma etapa fundamental de pré-processamento em clusterização. Justifique relacionando com o item da questão anterior.

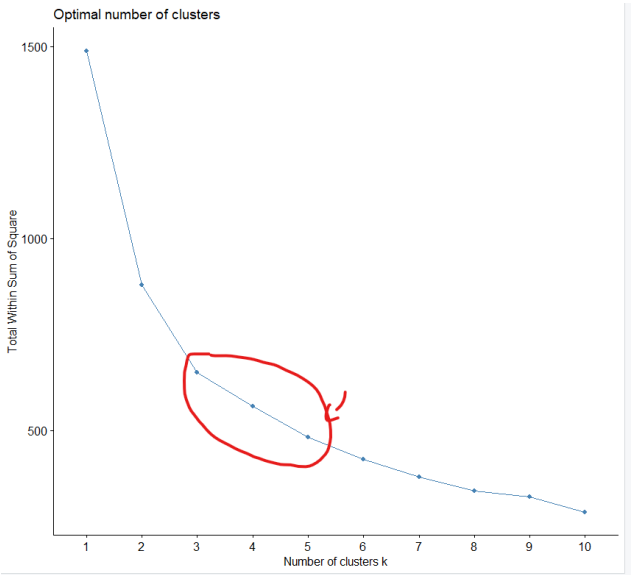
Resp: **O objetivo da normalização é colocar todos os valores das colunas numéricas em uma escala comum, sem distorcer as diferenças dos intervalos de valores. Em muitos casos é de extrema importância realizar a normalização dos dados, antes de aplicar a clusterização (Ex. Quando utilizamos a Distância de Manhattan, é importante realizar a normalização dos dados).**
Contudo, há alguns casos em que a clusterização fica melhor sem a normalização. Esse método é necessário apenas quando os parâmetros tiverem intervalos muito divergentes.

3. Aplique o algoritmo de K-Médias nos dados normalizados (função scale). Para tal você irá determinar o número de centróides que melhor atende o seu problema. Justifique a escolha (a justificativa pode ser empírica) e apresente os resultados.
4.

Resp: **A fim de definir o número de centróides, temos que aplicar alguma das técnicas de verificação do número ideal de clusters. Para o gráfico abaixo, utilizamos no R: "fviz_nbclust(Dados_scaled, kmeans, method = "wss") # Resultado entre (3k), (4k) ou (5K)? #" ("Wss" = Within sum squares) Uma das técnicas que podemos utilizar.**

Entendo que podríamos utilizar 3, 4 ou 5 para este caso. Contudo, utilizamos 5. E tivemos o seguinte resultado:

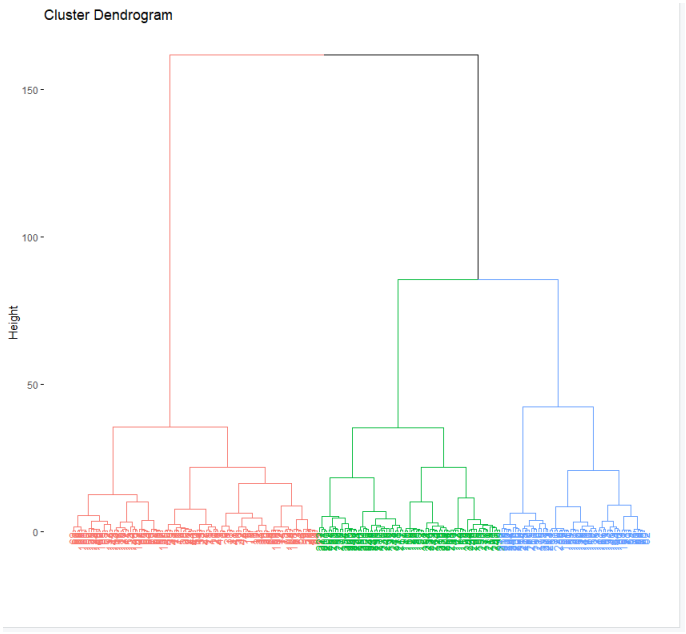
```
Within cluster sum of squares by cluster:
[1]  52.17368  61.69209 121.61635 167.02907  87.74273
 (between_SS / total_SS =  67.1 %)
```



5. Aplique o algoritmo de clusterização hierárquica nos dados normalizados (função scale).

Resp: **Feito no Rstudio.**

6. Mostre o dendograma da clusterização hierárquica. Quantos clusters são indicados pelo dendograma.
   Resp: **Feito no Rstudio. Pelo dendrograma abaixo, entendo que deveriam ser utilizados 3 clusters.**



7. Compare os dois algoritmos utilizados e escreva suas conclusões.

Resp: **A clusterização hierárquica - Agrupa objetos em clusters com uma estrutura de árvore. A vantagem dele é que não requer um número especificado de clusters, diferentemente de todos os outros clusters. E possui uma boa visualização através dos dendrogramas e facilita bastante a visualização e relação entre os clusters. Ainda, pode também ser utilizado para análise exploratória dos dados e analisar o número de cluster a serem utilizados.**

Cluster com o K Means - Necessário a definição dos centróides para representação dos cluster, e com isso possui uma certa sensibilidade em relação a escolha da quantidade de cluster. Por isso é importante realizar uma análise para definir o número ideal de centróides.

Dentre os 2, eu achei melhor para a visualização a clusterização hierárquica. Não apenas na produção deste trabalho, mas em meus estudos em geral durante a disciplina. A clusterização hierárquica possui uma melhor visualização na minha opinião, até mesmo para a definição do número de clusters. Contudo, entendo a necessidade de por exemplo utilizar um CLARA para base de dados com muitos grandes.