Antonio,

Nessa disciplina, aprendemos sobre segmentação e clusterização. Organizar dados de acordo com a similaridade estatística é uma tarefa extremamente importante para o dia-a-dia de um cientista de dados. Para nos auxiliar, vimos algumas aplicações com a linguagem R. Agora iremos validar nosso conhecimento.

* 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'. (AULA 3, 1:50 HS )
  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 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](https://www.wired.com/story/finding-lena-the-patron-saint-of-jpegs/)) é muito utilizada em estudos com processamento de imagens. Use o algoritmo CLARA (vide aula 7 – 1:24) – linguagem r – aula 6 ) nesta imagem e a represente com 3, 5 e 10 cores distintas. Escreva suas conclusões referentes a qualidade de representação da imagem (a informação é da imagem é perdida? Melhora com mais cores?)

Arquivo Lena

<https://media.wired.com/photos/5c523ece95d28a2cb5b76db1/191:100/w_1280,c_limit/Backchannel-Lena-Soderberg-FA.jpg>

usar como .PNG

https://upload.wikimedia.org/wikipedia/en/7/7d/Lenna\_%28test\_image%29.png

**Atenção : Para esta questão veja a aula5 – 1:30**

**Mulher com chapéu na cabeça

Descrição gerada automaticamente**

AULA 4 – 1:15HS (K-MEANS)

Sempre que usar o K-means, utilizar o comando set.send() para que o professor saiba qual agrupamento utilizamos.

Vide aula 5 – 30:00 (listas aleatórias)

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

3. Escolha uma base de dados para realizar esse projeto. Essa base de

(Vide aula 3 – https://archive.ics.uci.edu/)

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Base escolhida para o trabalho :

<https://archive.ics.uci.edu/dataset/186/wine+quality>

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).

Obs.: (Utilizar entre 6 e 10 variaveis. Usar dados numéricos. Não usar dados categóricos)

Obs.: Verificar se as grandezas da base estão na mesma escala (vide aula 5 – 35:00, arquivo do R da aula 4), caso negativo será necessário usar a função scale para ajustar as informações

antes

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Depois

Interface gráfica do usuário

Descrição gerada automaticamente

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.

4. Para essa questão utilize o pacote factoextra e a base escolhida na questão anterior: (factoextra – vide aula 6)

* + - 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.

Interface gráfica do usuário, Texto, Aplicativo, Word, Email

Descrição gerada automaticamente

Obs.: ( utilizar este artivo para responder esta questão )

* + - 1. A normalização dos dados é uma etapa fundamental de preprocessamento em clusterização. Justifique relacionando com o item de questão anterior.

Z-Score / MinMax Scaling -> Aula 3

Obs.: (vide aula 5 – 14:20 - Z-Score / MinMax Scaling)

* + - 1. 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. (aula 4)
      2. Aplique o algoritmo de clusterização hierárquica nos dados normalizados ( função scale).(aula 7)
      3. Mostre o dendograma da clusterização hierárquica. Quantos clusters são indicados pelo dendograma.
      4. Compare os dois algoritmos utilizados e escreva suas conclusões.

Antes de fazer sua entrega, reúna todos os arquivos relativos ao seu Projeto de Disciplina em um único arquivo no formato .zip e poste no Moodle. Utilize o seu nome para nomear o arquivo, identificando também a disciplina, como no exemplo: “nomedoaluno\_nomedadisciplina\_pd.zip”.