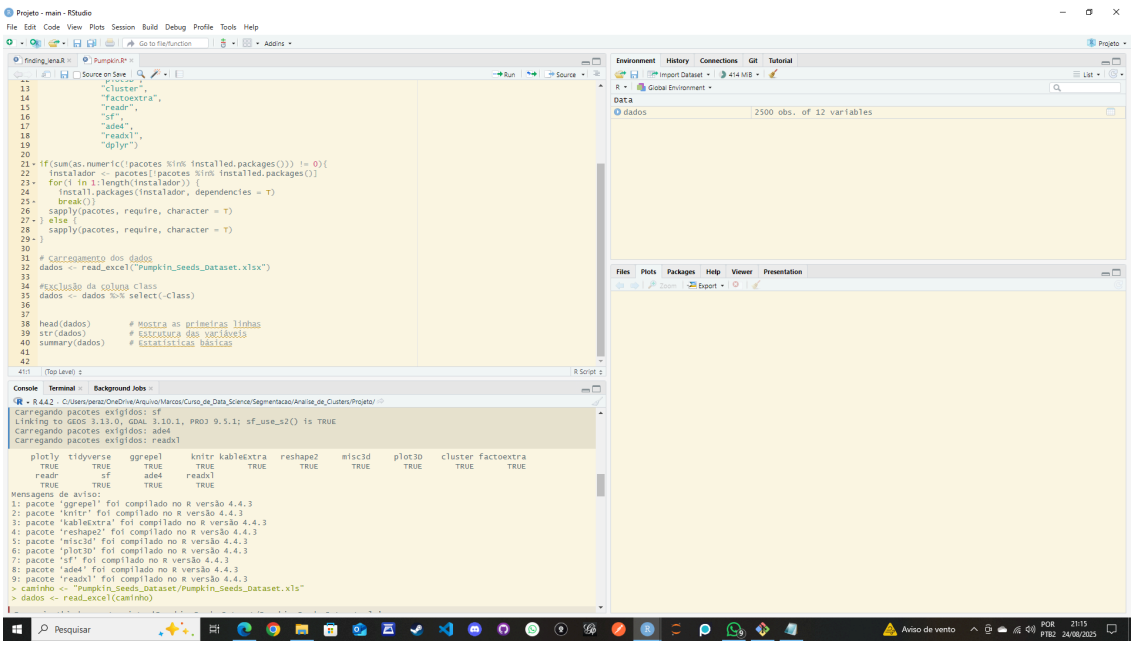
Instituto Infnet

Análise de Clusters

Professor: Gesiel Rios Lopes
Aluno: Marcos Perazo Viana
25 de agosto de 2025

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'.



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

Imagem original

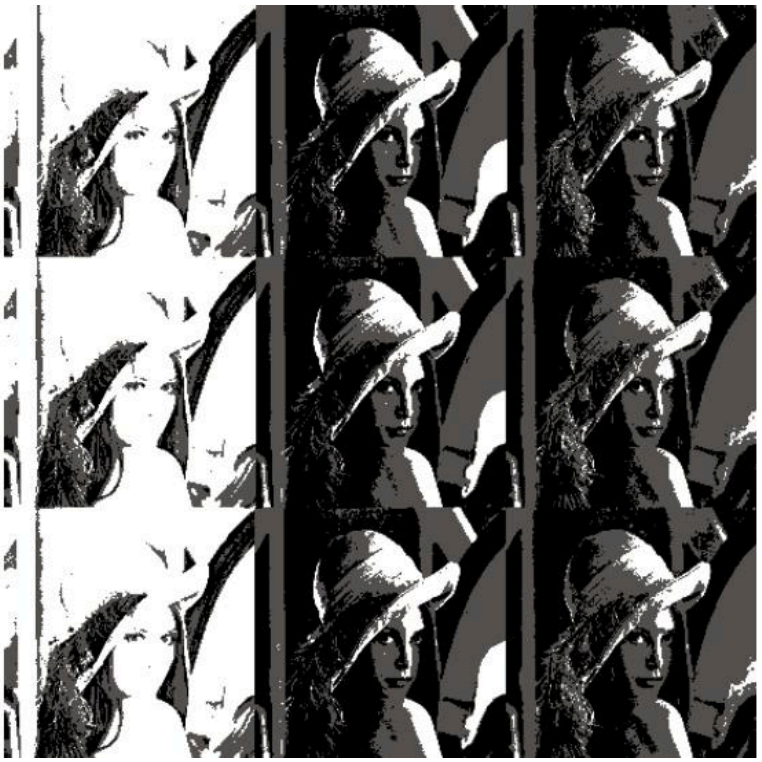


Imagem com três cores

imagem com cinco cores

Imagem com 10 cores

A técnica de clusterização das tonalidades dos pixels por meio do algoritmo CLARA mostrou-se eficaz para realizar quantização de cores na imagem da

Lena. Ao agrupar os pixels em torno de medoids representativos, conseguimos reconstruir a imagem com 3, 5 e 10 cores distintas, reduzindo significativamente a complexidade cromática.
Com três cores, a imagem apresenta forte perda de detalhes. As áreas de sombra, contorno e textura são simplificadas ao ponto de comprometer a identificação precisa de elementos faciais. A informação visual é bastante reduzida, e a imagem se aproxima de uma arte abstrata.
Com cinco cores, há uma melhora perceptível na definição. Embora ainda simplificada, a imagem começa a recuperar contornos e áreas de contraste. É possível reconhecer melhor os traços do rosto e os elementos principais, embora nuances sutis ainda estejam ausentes.
Com dez cores, a representação se aproxima da imagem original. As transições de cor são mais suaves, e os detalhes faciais, como olhos, boca e cabelo, ganham nitidez. A perda de informação é mínima, e a imagem mantém boa fidelidade visual.
A conclusão geral é que a qualidade de representação da imagem melhora proporcionalmente ao número de cores utilizadas. Com poucos clusters, a simplificação é extrema e compromete a legibilidade. À medida que aumentamos o número de clusters, a imagem recupera detalhes e profundidade, demonstrando que há um trade-off entre compressão e fidelidade visual.

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.

A base de dados escolhida se encontra no endereço Pumpkin Seeds Dataset. Ela foi criada para classificação de sementes com base em características morfológicas extraídas de imagens. Isso tem aplicações diretas em: Agricultura de precisão, estudos botânicos e em processos industriais de seleção de sementes.

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.

À medida que o número de dimensões aumenta, todas as distâncias tendem a se tornar semelhantes. A diferença entre o ponto mais próximo e o mais distante de um centróide se torna insignificante, dificultando a separação clara entre clusters.
A distância euclidiana considera todas as dimensões igualmente. Se houver variáveis ruidosas ou irrelevantes, elas podem distorcer a medida de distância e comprometer a formação dos agrupamentos.
Mesmo com normalização, em alta dimensionalidade, pequenas variações em muitas dimensões podem somar e gerar distâncias artificiais.
A distância de Mahalanobis seria uma boa opção em relação a K-Means, pois leva em conta a correlação entre variáveis. Também, ajusta a escala e orientação dos dados, tornando a medida mais precisa. E é excelente para dados com covariância significativa.

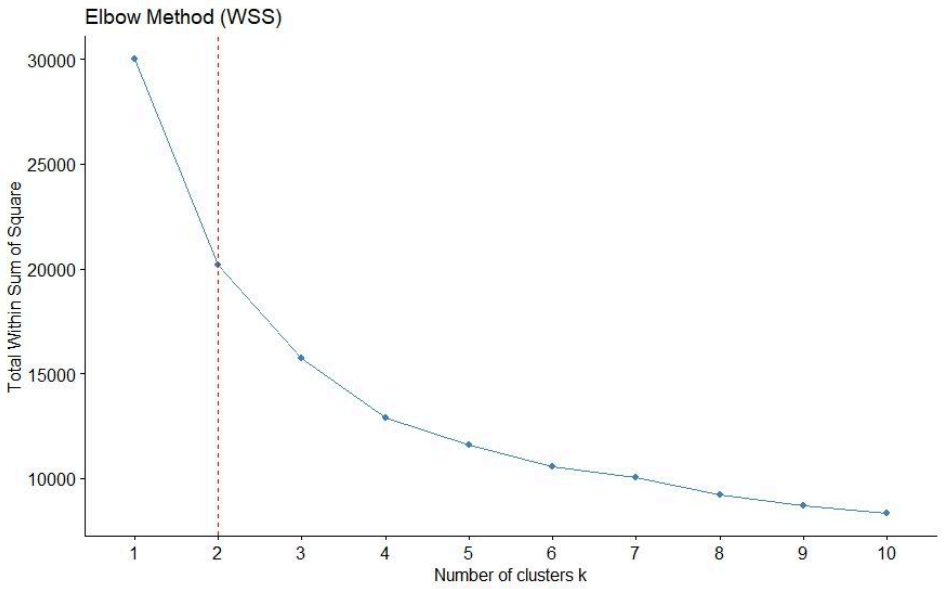
2 - A normalização dos dados é uma etapa fundamental de pré-processamento em clusterização. Justifique relacionando com o item da questão anterior.

A distância euclidiana é sensível à escala das variáveis. Se uma variável tem valores muito maiores que as outras. Por exemplo, levando em consideração a base de dados que será utilizada, se a área varia de 47939 a 136574, enquanto o perímetro varia de 868,5 a 1559,5, a área domina o cálculo da distância, distorcendo os agrupamentos.
Em alta dimensionalidade, esse problema se agrava, pois cada dimensão adiciona mais “peso” à distância total. Variáveis com escalas maiores influenciam desproporcionalmente a posição dos centróides. Isso pode levar o algoritmo a formar clusters baseados em variáveis irrelevantes ou ruidosas.
A normalização, como z-score ou min-max scaling, transforma todas as variáveis para uma mesma escala, geralmente com média 0 e desvio padrão 1 ou entre 0 e 1. Isso garante que todas as variáveis contribuam equivalentemente para o cálculo da distância. Garante também que o algoritmo de K-Médias possa identificar padrões reais nos dados, e não apenas diferenças de magnitude. E a performance do modelo melhore,

especialmente em espaços de alta dimensão, onde a concentração de distâncias é um problema.

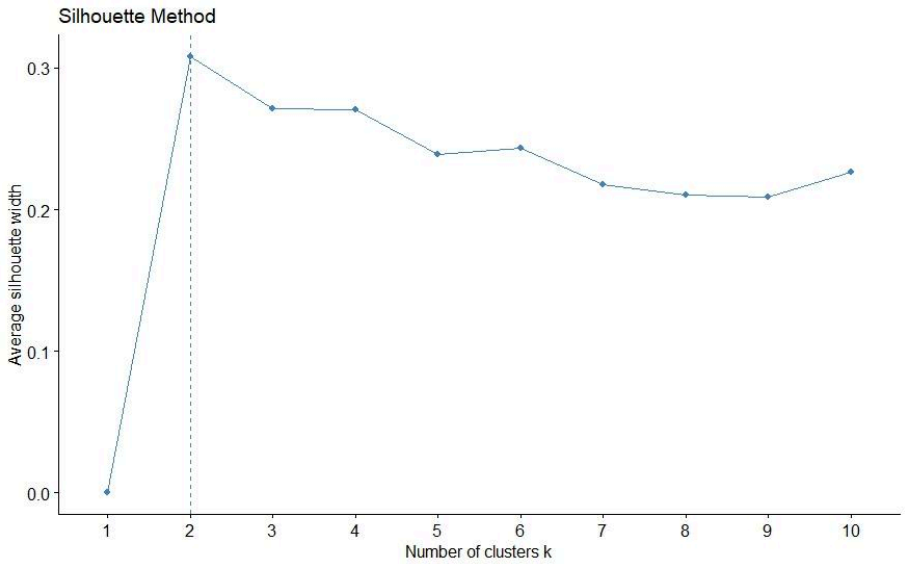
3 - Aplique o algoritmo de K-Médias nos dados normalizados (função scale). Para tal você irá determinar o número de centróides que melhor atende o seu problema. Justifique a escolha (a justificativa pode ser empírica) e apresente os resultados.

Utilizando o método do Cotovelo, obtivemos uma curva suave, o que torna difícil a leitura da curva, mas o vértice k=2 parece ter uma angulação maior em relação aos outros vértices.



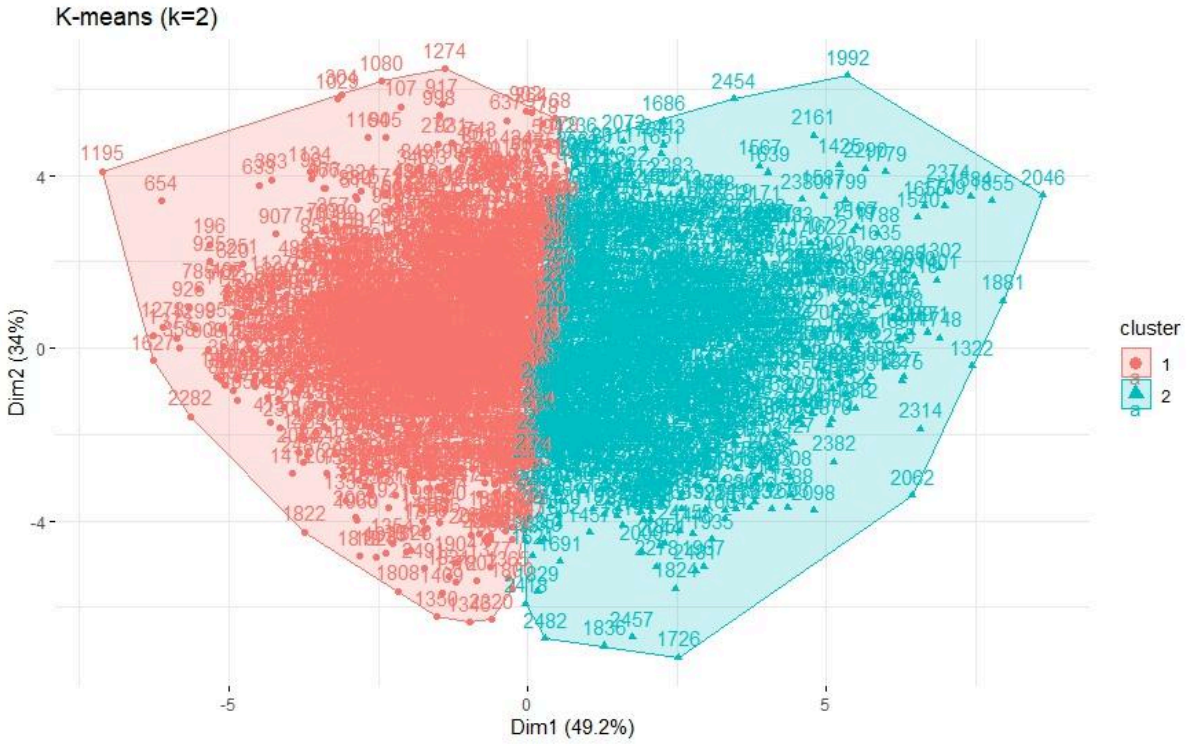
Método do Cotovelo

Utilizando o método da Silhueta obtivemos o pico em k=2, mas existem dois vértices que também devem ser analisados, em k=4 e k=6.



Método da Silhueta

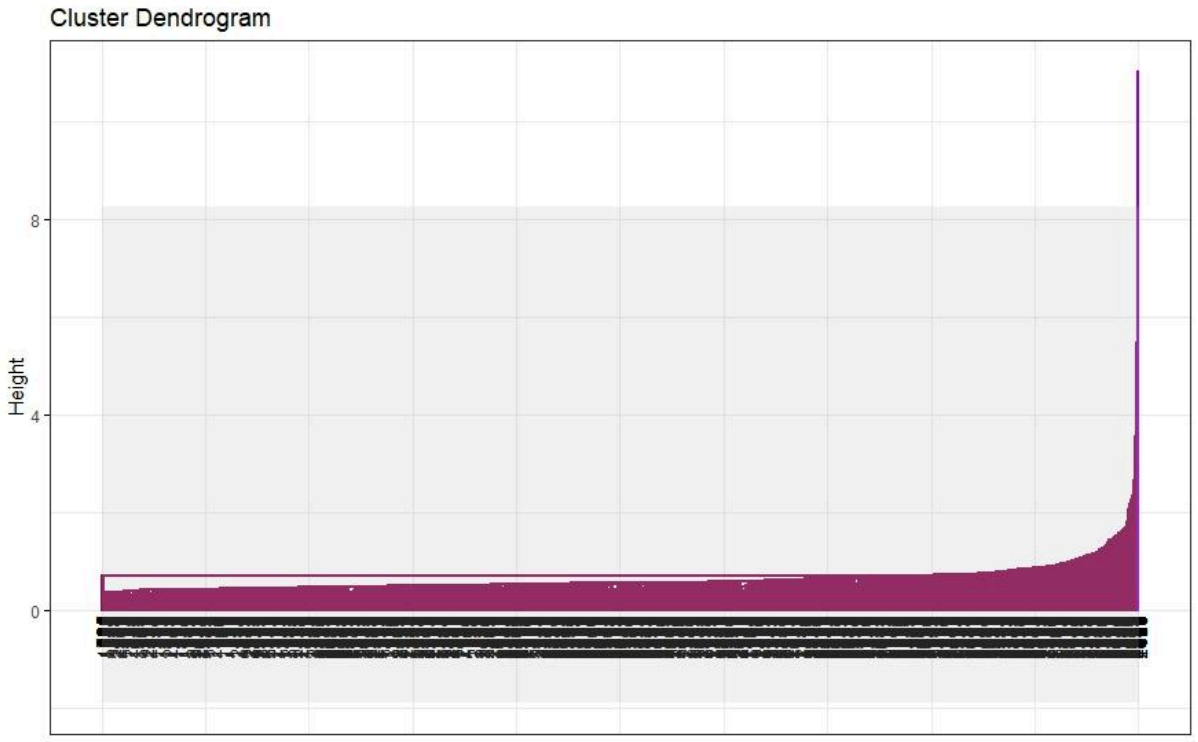
Segue o gráfico obtido após a aplicação do algoritmo K-Médias com k=2:



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

Algoritmo aplicado no arquivo Pumpkin.r

5 - Mostre o dendograma da clusterização hierárquica. Quantos clusters são indicados pelo dendograma.



Devido a grande quantidade de dados a legibilidade do dendograma se torna um desafio. Houve uma superlotação no eixo horizontal, pois cada linha representa uma observação, então com milhares de pontos, os rótulos se sobrepõem ou desaparecem. O número de junções cresce exponencialmente, tornando difícil identificar padrões visuais ou cortes significativos. As distâncias ficaram muito próximas, com dados normalizados e muitas variáveis, os agrupamentos podem parecer indistintos, dificultando a interpretação.

O dendograma revela que houve grande concentração de junções em alturas baixas, isso indica que muitos dados são bastante semelhantes entre si, formando grupos compactos. A base parece ter uma estrutura interna bem definida, com alta similaridade entre várias observações.

Quando os ramos começam a se unir em alturas maiores, isso sugere que os grupos que estavam bem separados estão sendo forçados a se unir. Esse ponto é ideal para cortar o dendrograma e definir o número de clusters.

Visualmente, há uma divisão mais clara em dois grandes blocos, o que reforça a escolha de k=2 feita anteriormente. Isso indica que a base pode ser segmentada em dois perfis principais de sementes.

6 - Compare os dois algoritmos utilizados e escreva suas conclusões.

A aplicação dos algoritmos de K-Means e Clusterização Hierárquica à base de dados morfológicos de sementes de abóbora permitiu uma avaliação clara da capacidade de agrupamento dos métodos, bem como da estrutura latente presente nos dados. Ambos os algoritmos foram utilizados após a normalização das variáveis, etapa fundamental para garantir que todas as dimensões contribuíssem de forma equitativa na medida de distância.

O algoritmo K-Means, de natureza particional, requer como entrada o número de clusters desejado. Para determinar esse valor, foram utilizados os métodos do cotovelo (WSS) e da silhueta, ambos indicando que a divisão ideal seria em dois grupos. A execução do K-Means com dois clusters resultou em agrupamentos bem definidos, com centróides representativos e separação visual clara entre os grupos. A eficiência computacional do K-Means, aliada à sua simplicidade, o torna especialmente adequado para bases estruturadas e de grande volume, como a utilizada neste estudo.

Por outro lado, a Clusterização Hierárquica, de natureza aglomerativa, foi aplicada utilizando o método de encadeamento single linkage, que considera a menor distância entre elementos de diferentes grupos para realizar a junção. Esse método permitiu a construção de um dendrograma que revelou a estrutura de aglomeração progressiva dos dados, evidenciando também a formação de dois agrupamentos principais. Embora mais sensível à presença de ruídos e menos escalável em grandes bases, a clusterização hierárquica oferece uma perspectiva complementar ao K-Means, permitindo a análise da hierarquia entre os agrupamentos e a visualização de diferentes níveis de granularidade.

Ambos os algoritmos convergem para a formação de dois grupos distintos, sugerindo que há uma separação natural entre os tipos de sementes com base nas características morfológicas analisadas. O K-Means demonstrou maior eficiência e clareza na segmentação, sendo mais indicado para aplicações práticas de classificação automática. Já a Clusterização Hierárquica mostrou-se valiosa na etapa exploratória, contribuindo para o entendimento da estrutura interna dos dados.

Em síntese, a utilização combinada dos dois métodos proporcionou uma análise robusta e multifacetada da base de dados, reforçando a importância da escolha criteriosa de algoritmos conforme o objetivo da análise e a natureza dos dados envolvidos.