



Desvendando Padrões Musicais: Uma Análise Não Supervisionada de Dados de Música (1950-2019)

Este projeto explora o universo da música através de técnicas de aprendizado de máquina não supervisionado, revelando estruturas e tendências ocultas em um vasto conjunto de dados musicais.

Agenda do Projeto



Introdução e Objetivos

Definição do problema e metas do projeto de ML não supervisionado.



Metodologia

Carga, pré-processamento e análise exploratória do conjunto de dados.



Algoritmos Aplicados

Detalhamento das técnicas de agrupamento, redução de dimensionalidade e detecção de anomalias.



Resultados e Análise

Interpretação das visualizações e métricas, discutindo a qualidade das descobertas.



Implicações e Conclusões

Discussão sobre a utilidade prática do projeto e próximos passos.

O Problema: Explorando Estruturas Ocultas em Dados Musicais

Nosso objetivo principal é desvendar padrões e características intrínsecas no vasto universo musical, sem a necessidade de rótulos pré-definidos. O aprendizado não supervisionado permite-nos:



Agrupamento de Músicas

Identificar gêneros, estilos ou eras musicais emergentes.



Redução de Dimensionalidade

Simplificar dados complexos para visualização e análise.



Detecção de Anomalias

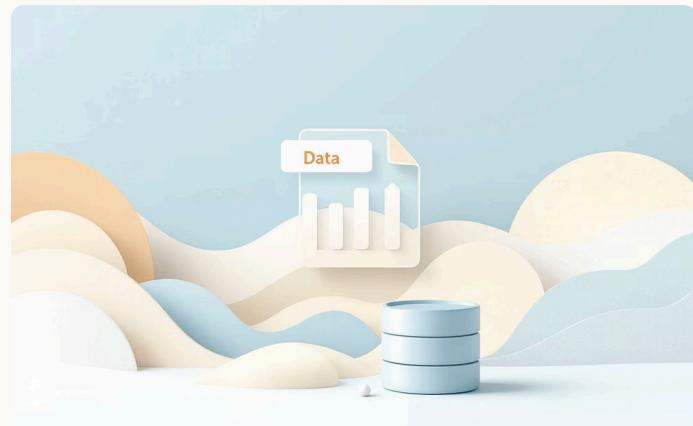
Encontrar músicas ou tendências atípicas que se destacam.



Conjunto de Dados e Pré-processamento

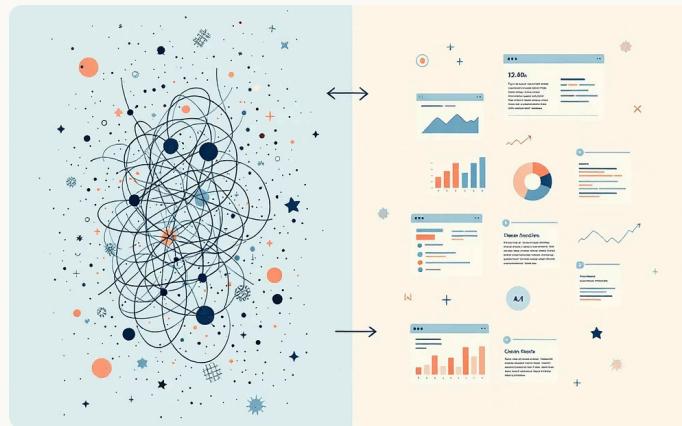
Utilizamos o "Music Dataset: 1950 to 2019 - Lyrics and Metadata", contendo informações detalhadas sobre músicas ao longo de sete décadas.

Etapas Realizadas:



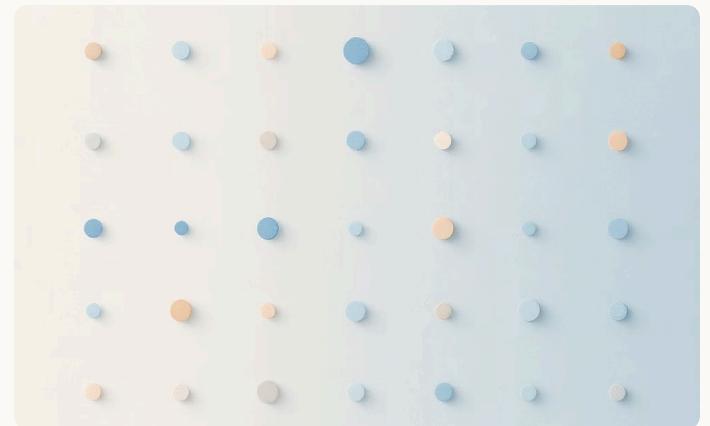
Carregamento dos Dados

Leitura do dataset "`tcc_ceds_music.csv`" (1950–2019) via Kaggle API e seleção de **22 variáveis numéricas** relevantes (ex: *danceability*, *valence*, *energy*).



Limpeza e Normalização

Conversão forçada para tipo numérico (*coerce*) e tratamento de valores ausentes com **imputação pela média**.



Padronização das Variáveis

Aplicação do **StandardScaler** para escalonamento (média = 0, desvio padrão = 1), garantindo comparação equitativa entre features. Dados finais: matriz padronizada de dimensão (*n_amostras*, 22).

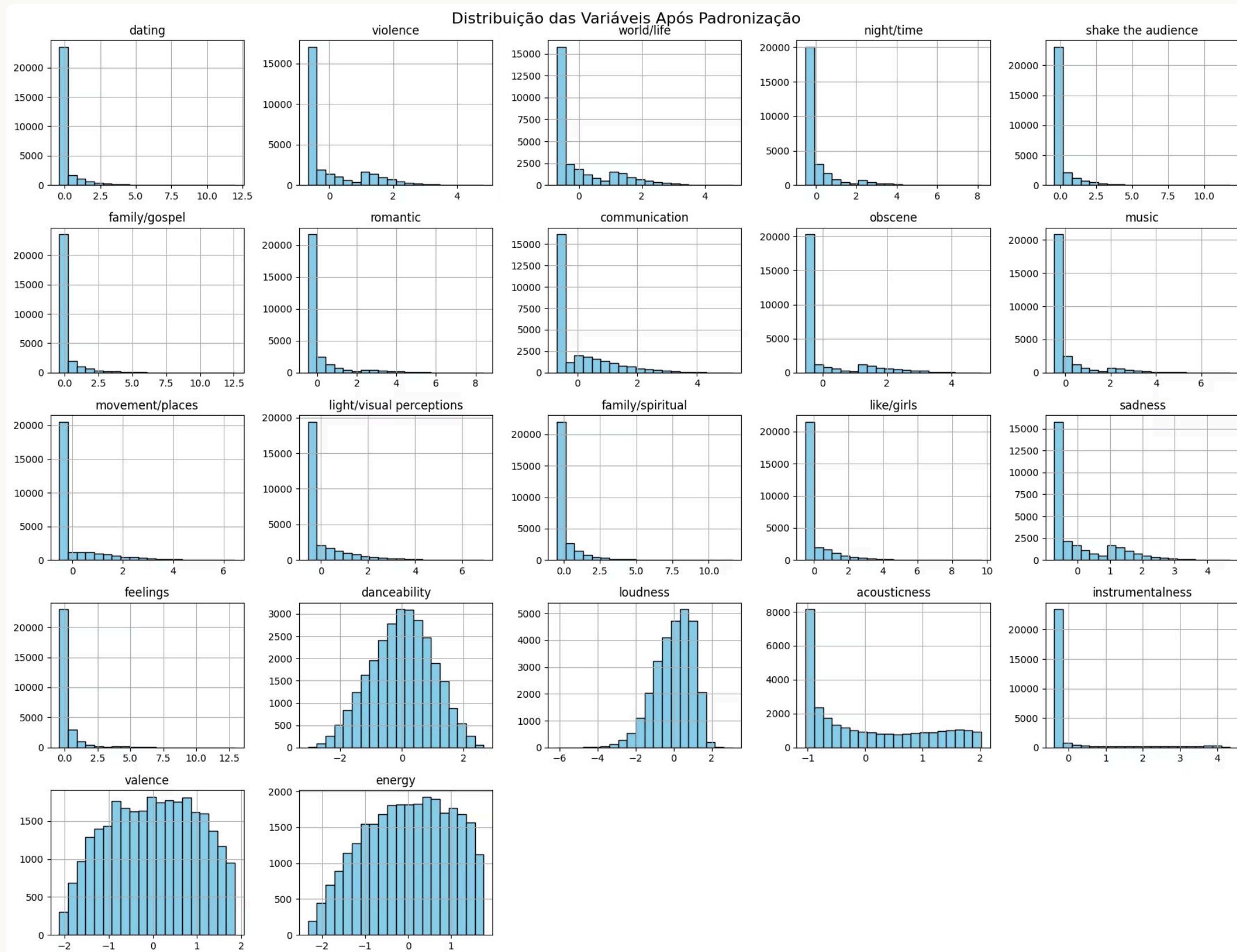
Exemplo de features: *violence* (letras violentas), *acousticness* (acústica), *romantic* (temas românticos).

O objetivo principal dessas etapas é preparar uma base de dados robusta e consistente para a aplicação de algoritmos de clustering.

Conjunto de Dados e Pré-processamento

Análise Exploratória (EDA):

- Visualização das Distribuições das Variáveis Padronizadas.

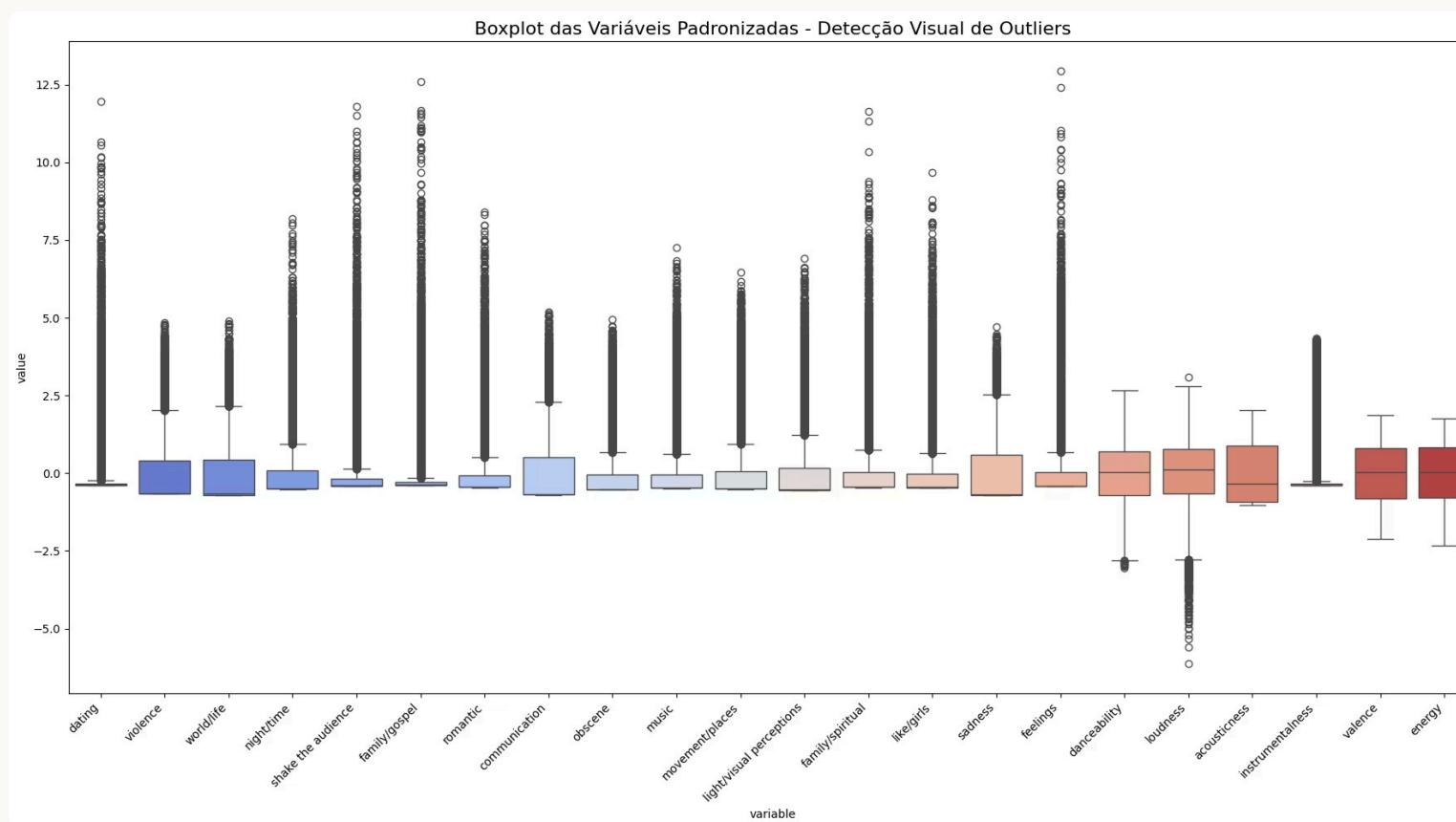


-

Conjunto de Dados e Pré-processamento

Análise Exploratória (EDA):

- Detecção de Outliers com Boxplot e IQR: Identificação de anomalias nos dados.



Interpretação Concisa dos Outliers:

- Muitos outliers** (ex: family/gospel, instrumentalness):
 - Refletem a **natureza polarizada** da música (ou é gospel ou não é; instrumental ou vocal).
 - São **subgêneros legítimos**, não ruído.
- Poucos/zero outliers** (ex: valence, energy):
 - Mostram **padrões industriais** (ex: músicas pop mantêm faixas similares de "energia").

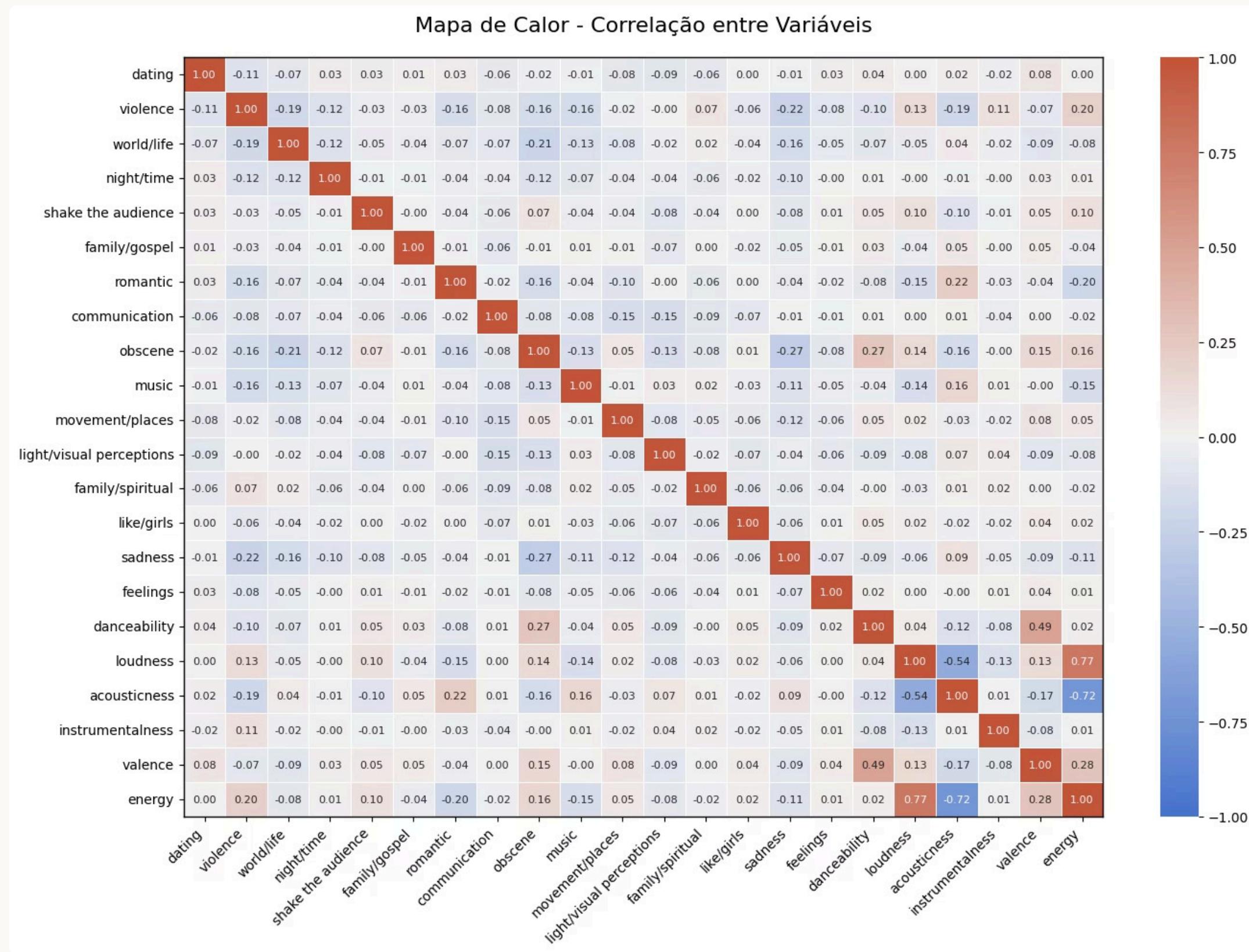
Ação sugerida:

Manter outliers para preservar diversidade musical, mas normalizar para análises estatísticas.

Conjunto de Dados e Pré-processamento

Análise Exploratória (EDA):

- Análise de Correlação com Heatmap: Compreensão das relações entre as variáveis.



Análise de Componentes Principais (PCA)

Objetivo do PCA:

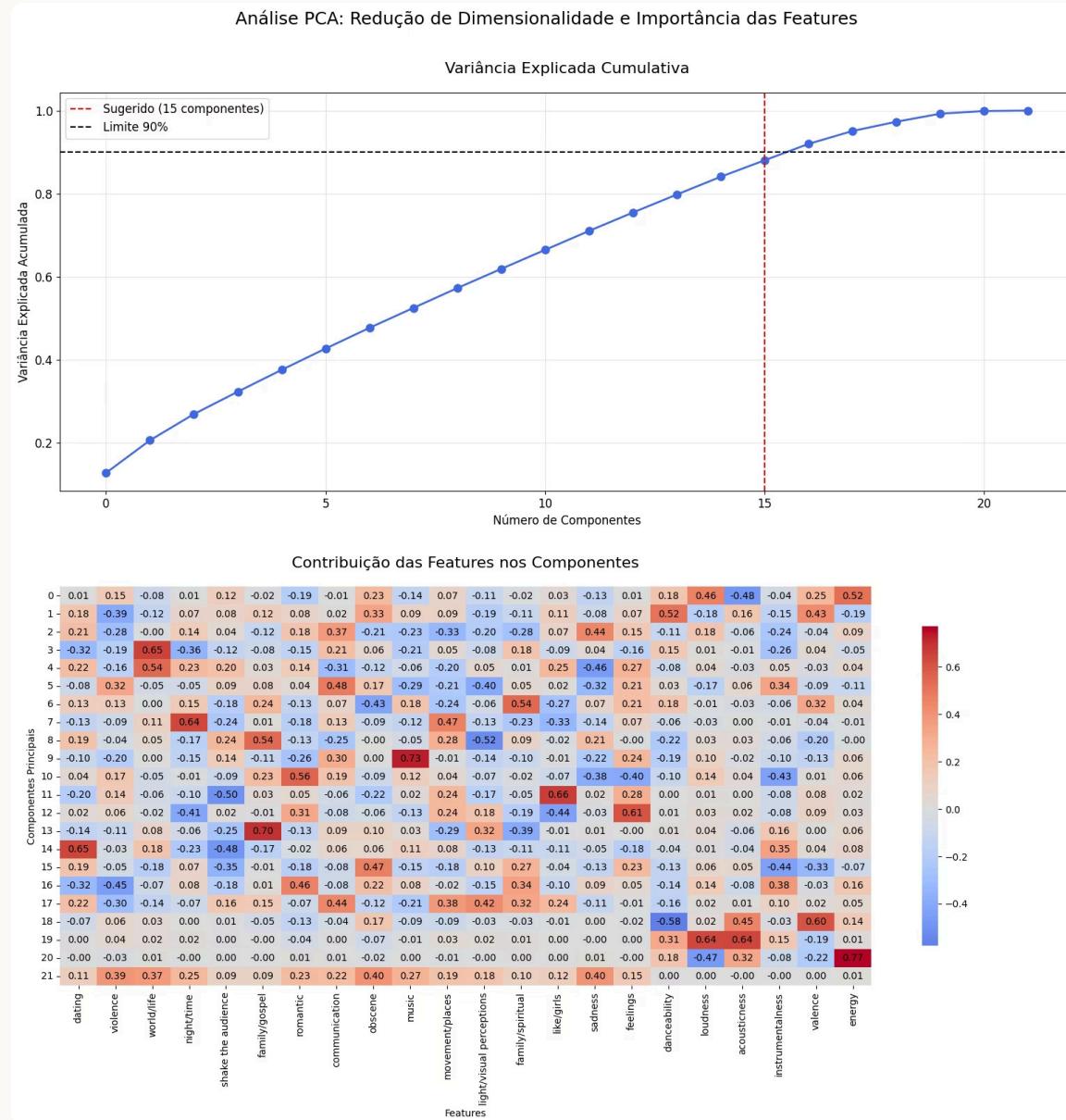
- Transformar variáveis correlacionadas em componentes independentes, reduzindo dimensões **mantendo a variância principal**, facilitando a visualização de dados de alta dimensão.

Variância Explicada:

- 17 componentes = 95% da informação (ideal para análise).
- 15 componentes = ~89% (opção conservadora).

🔥 Principais Componentes PC1 + PC2 = 20.6% da variância (base para visualização 2D).

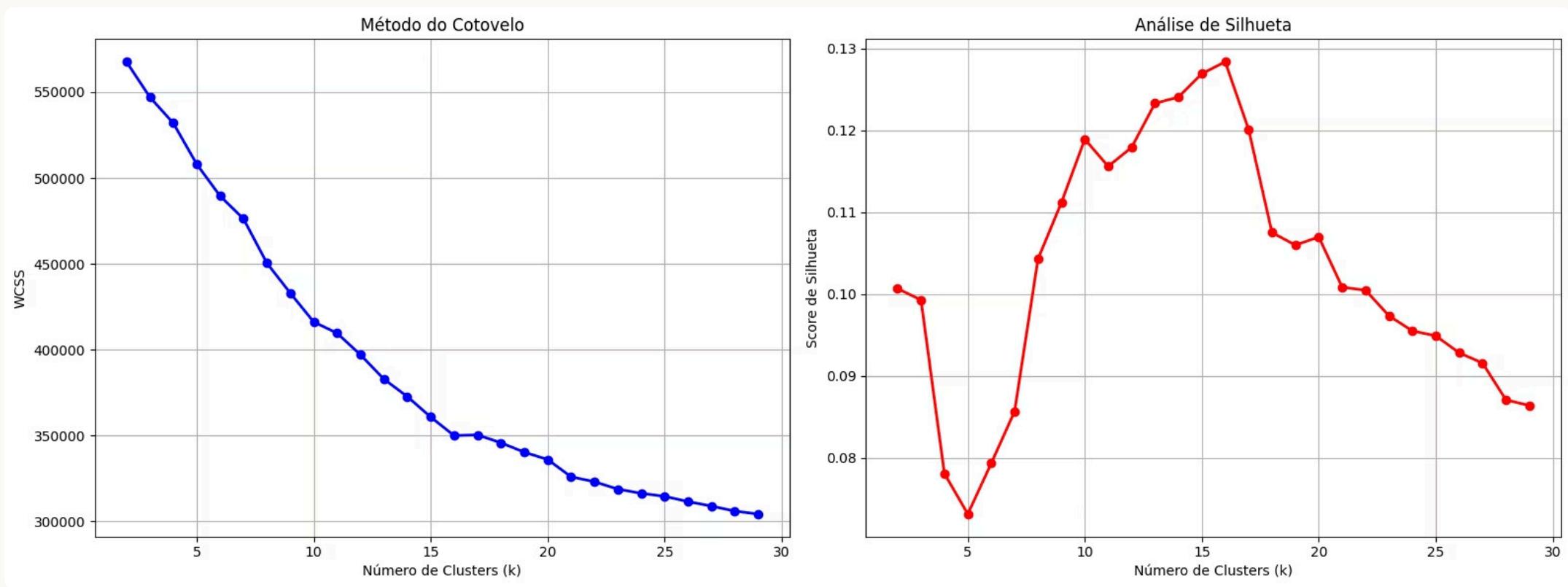
Heatmap revela features com $|valor| > 0.5$ definem cada componente.



Avaliação do Número Ideal de Clusters (K)

Visualização: Método do Cotovelo x Silhueta

- A análise visual do Método do Cotovelo e da Silhueta nos ajuda a determinar o número ideal de componentes principais e agrupamentos (k).



Algoritmos de Agrupamento: K-Means e Hierárquico

Aplicamos o K-Means e o Agrupamento Hierárquico para identificar grupos naturais dentro do nosso dataset musical.



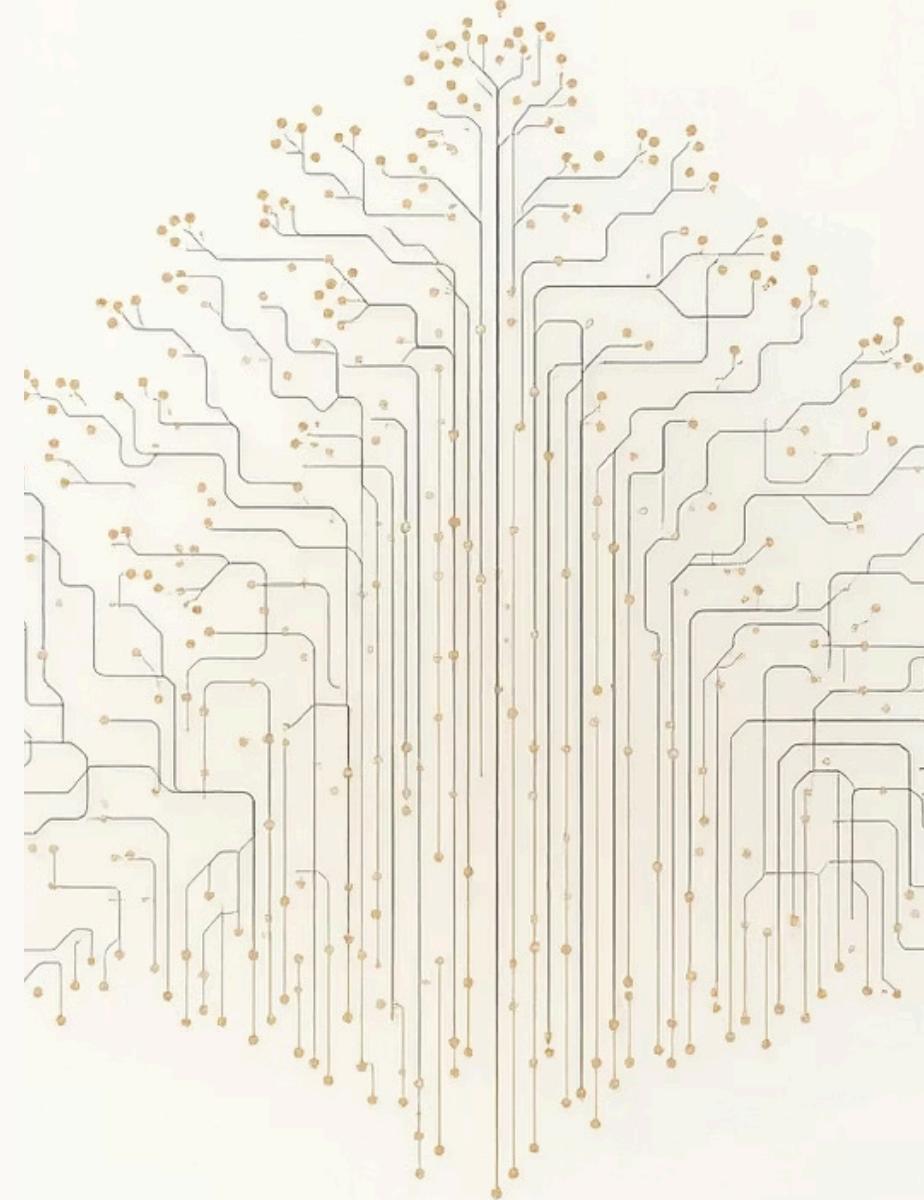
K-Means

Agrupamento baseado na similaridade de características, com experimentação de K variando entre 3 e 8, e com e sem PCA.



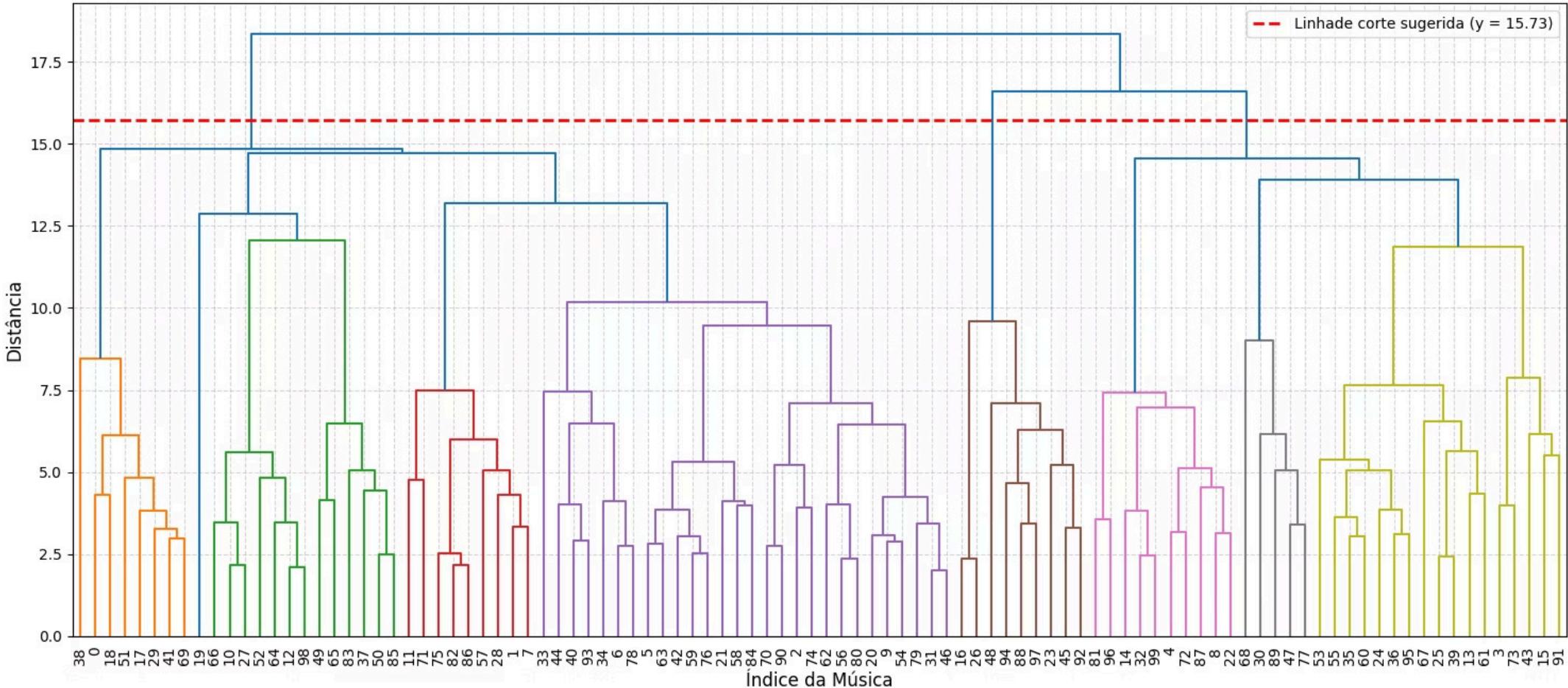
Agrupamento Hierárquico

Construção de uma hierarquia de clusters, visualizada através de um Dendrograma.

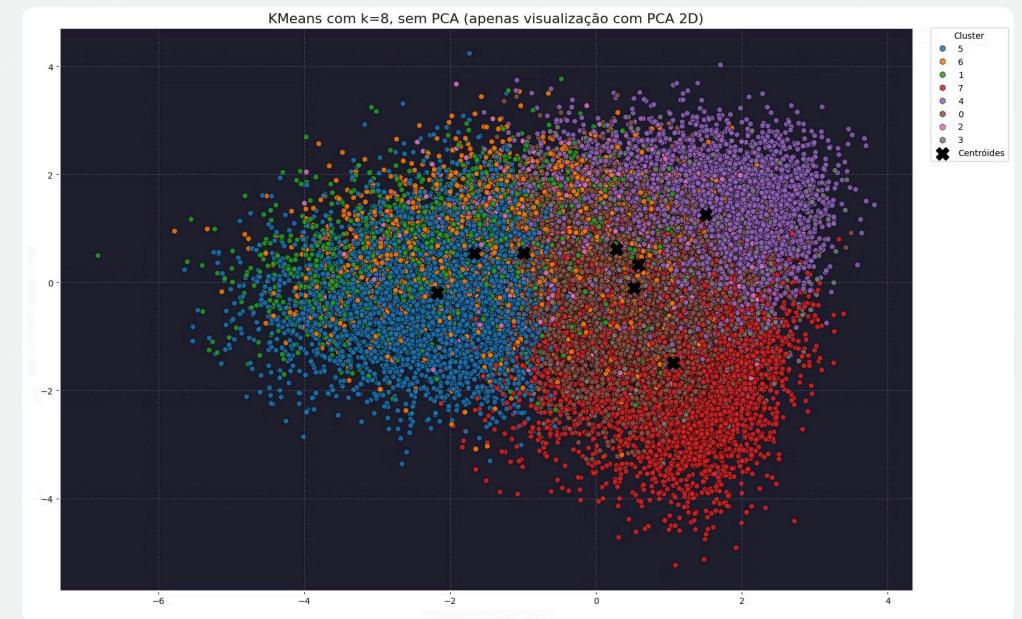
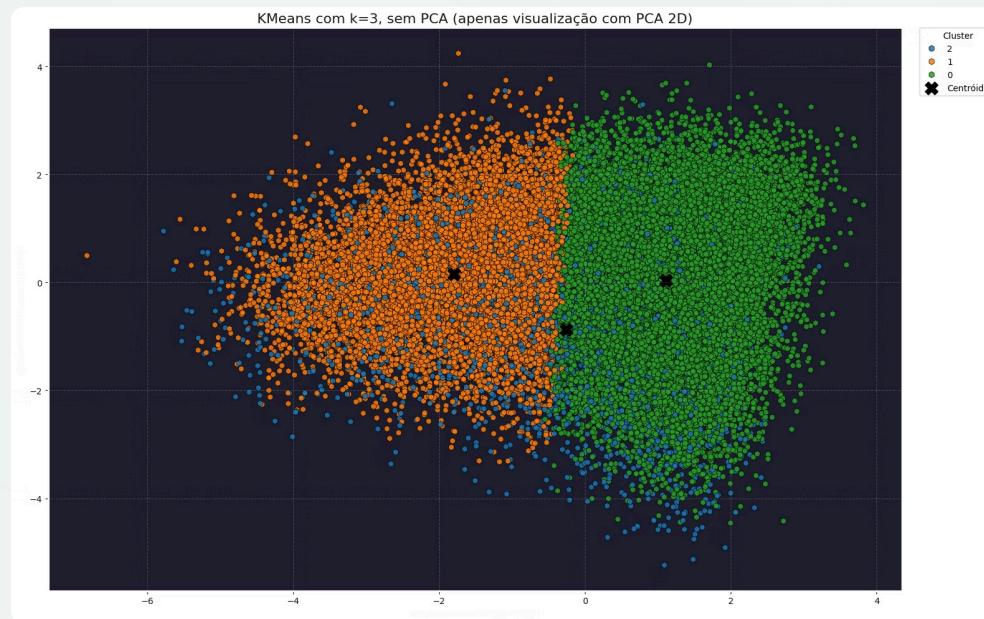


Dendrograma

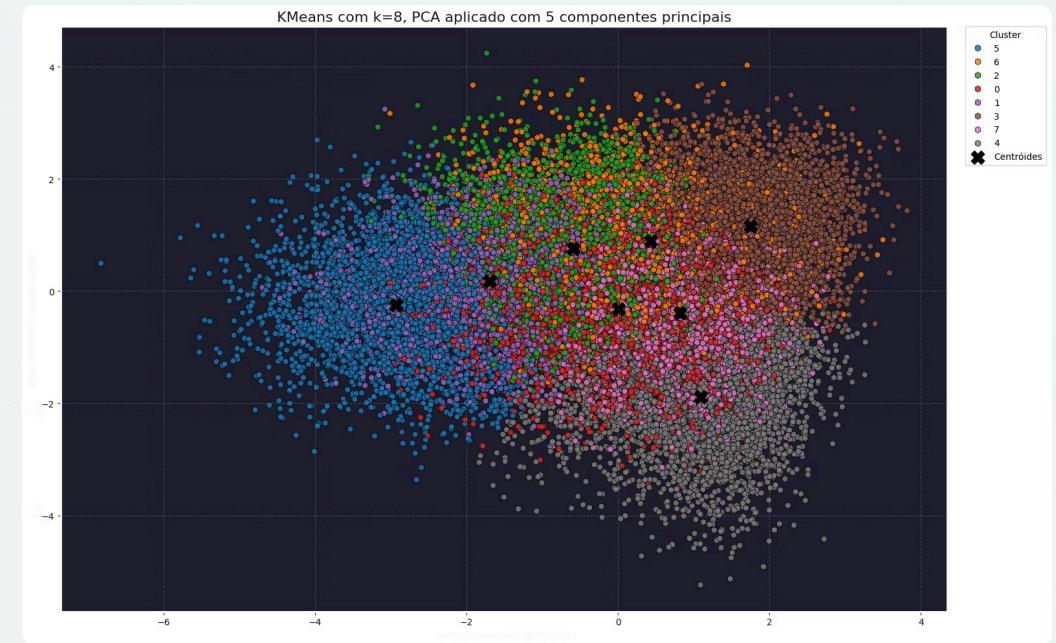
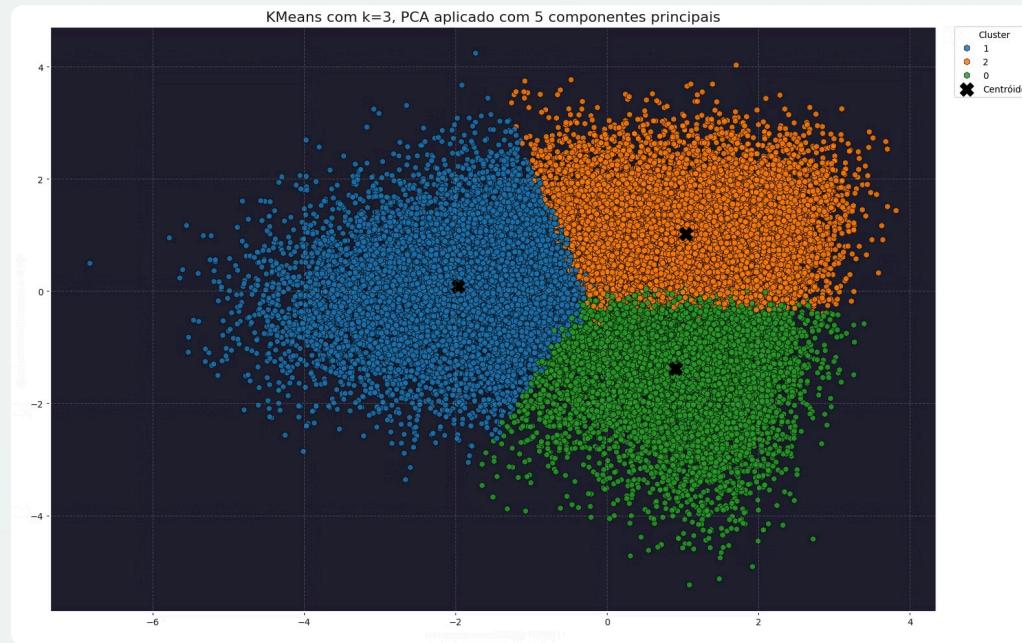
Dendrograma - Hierarchical Clustering



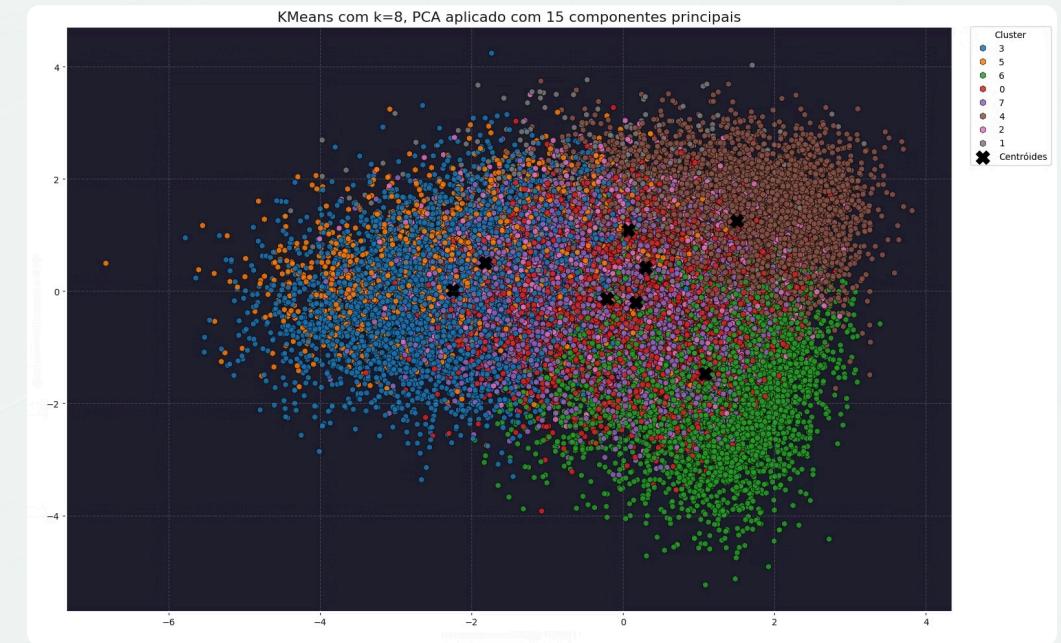
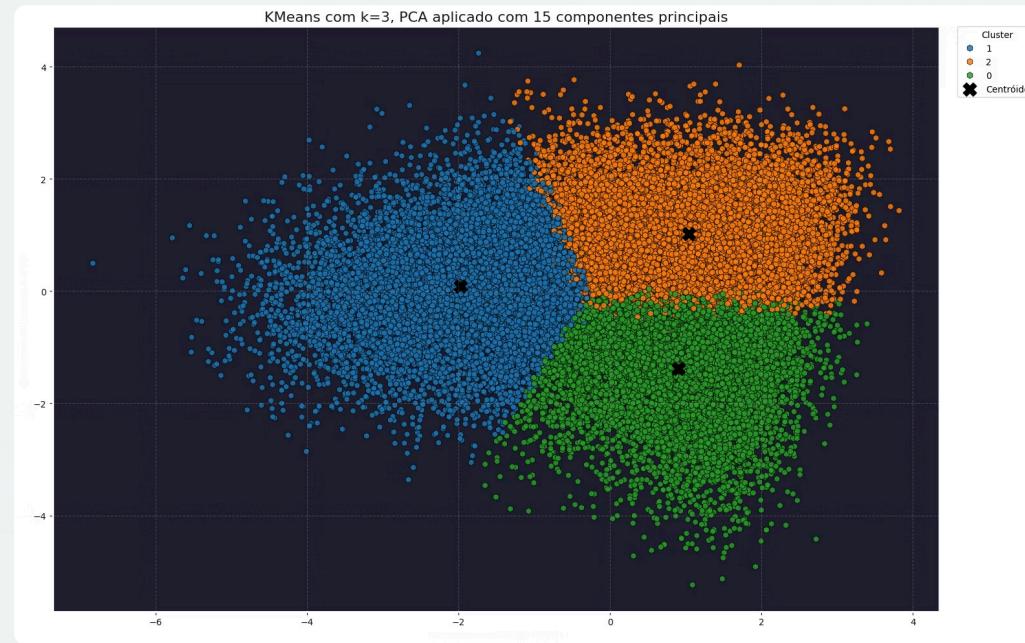
K-means aplicado às features originais (sem redução dimensional) e Análise de K=3 e K=8



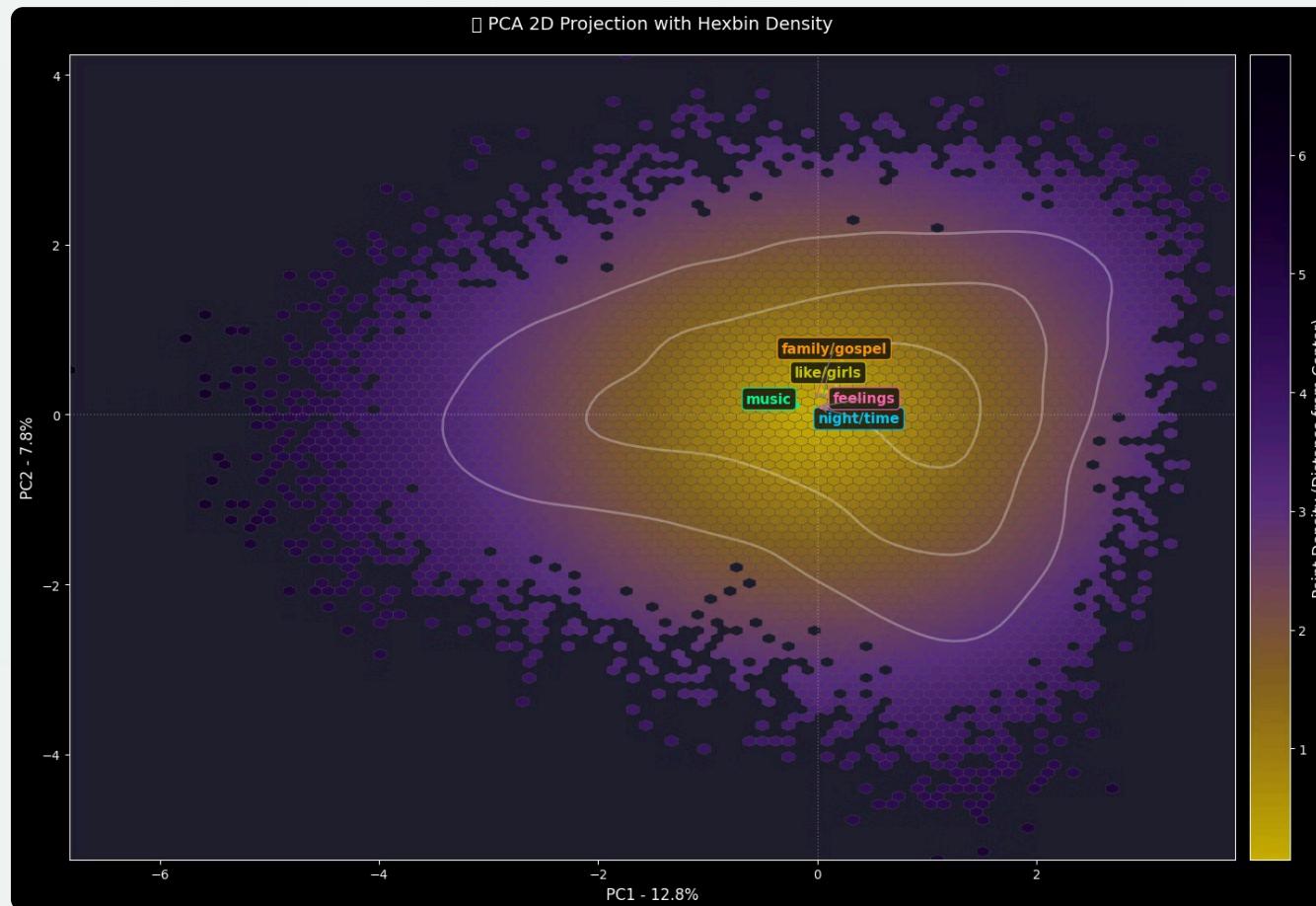
Clusterização com K-means: PCA (5 componentes) e Análise de K=3 e K=8



Clusterização com K-means: PCA (15 componentes) e Análise de K=3 e K=8



Visualização 2D com PCA: Intensidade Baseada em Componentes 17D

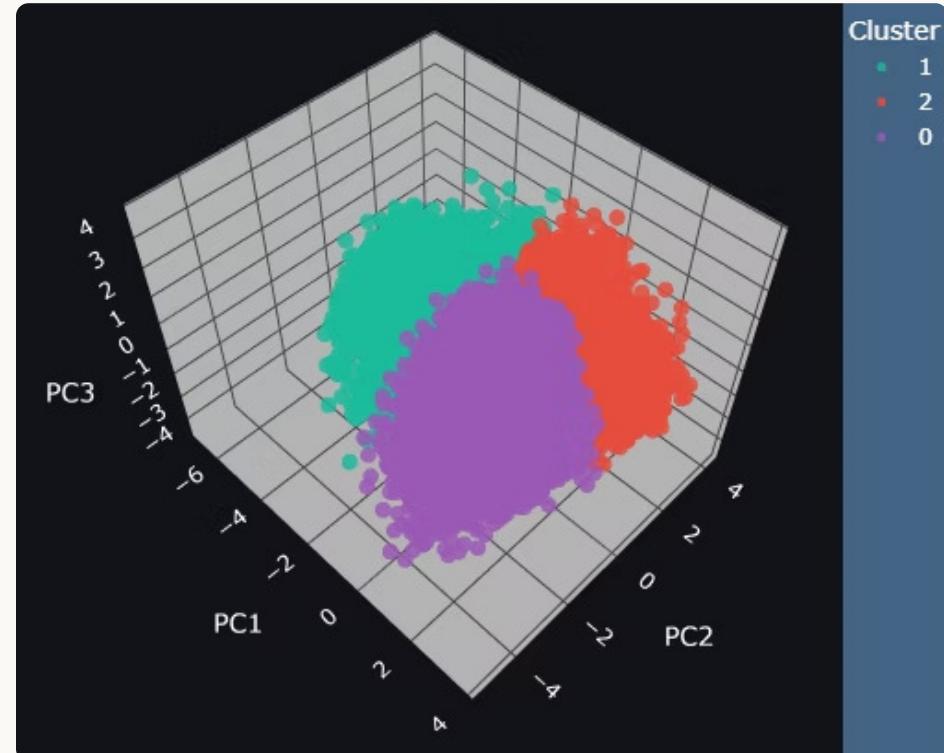
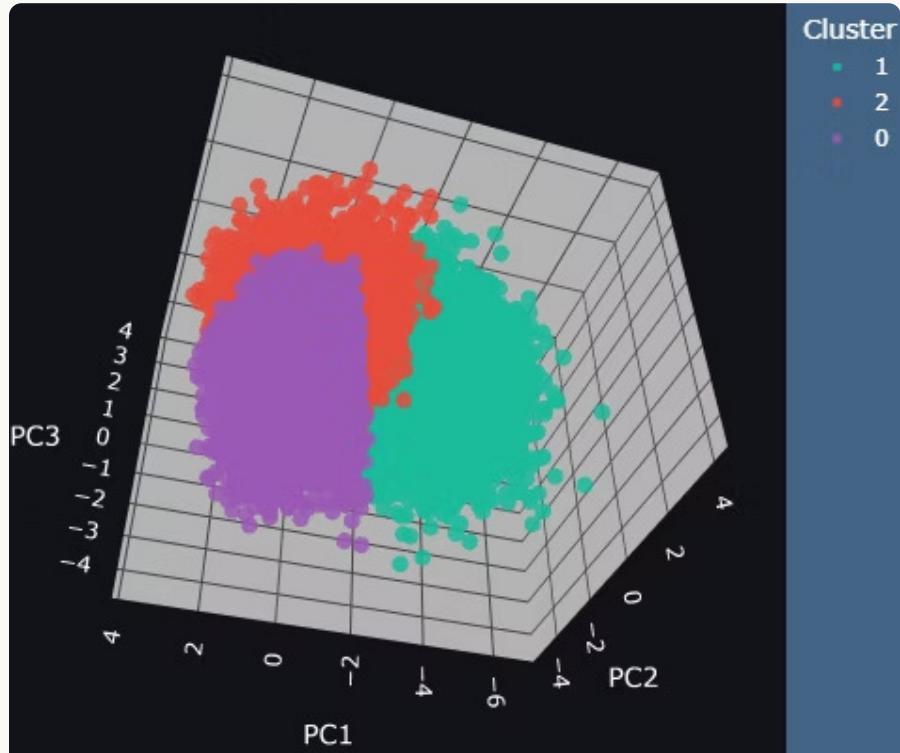


As features são anotadas no gráfico conforme sua contribuição aos eixos PC1/PC2 (tamanho/direção das setas), destacando as que mais diferenciam as músicas no espaço 2D."

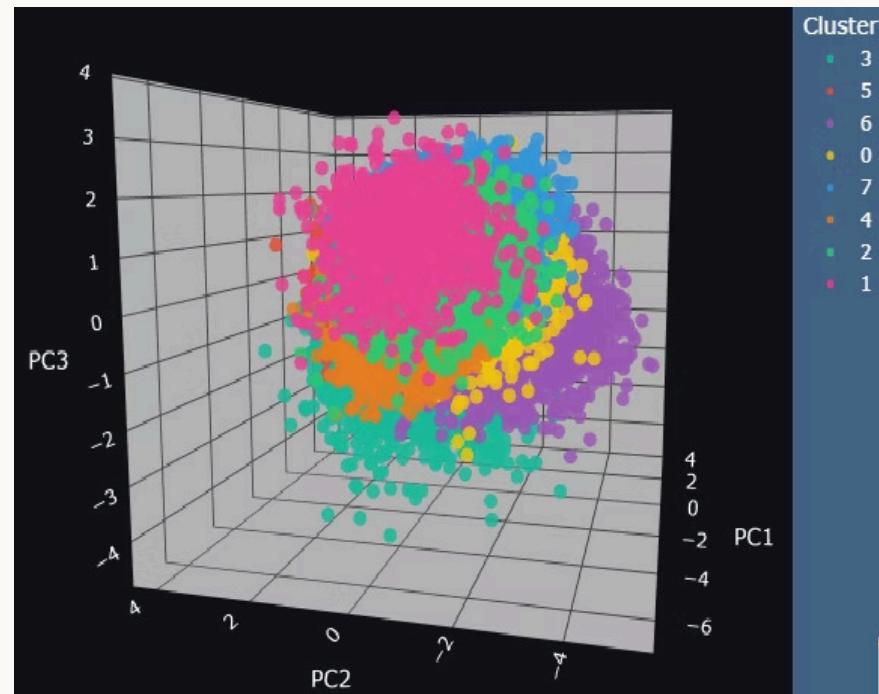
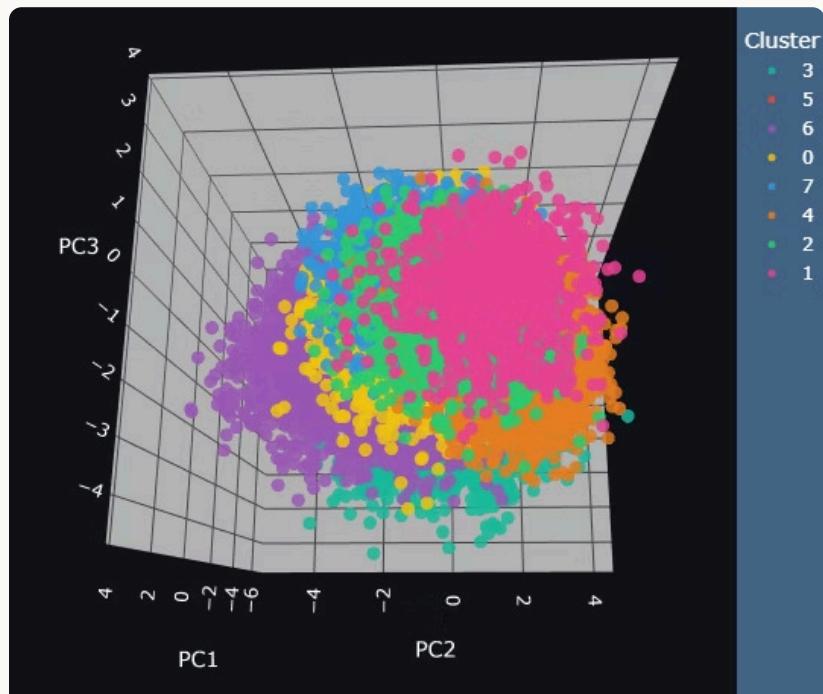
Por que aparecem?

- São as **5 features com maior magnitude nos *loadings*** (coeficientes) do PCA, calculados a partir das 17 componentes originais.
- Sua posição e direção refletem **como cada variável 'puxa' a distribuição dos dados** (ex: *family/gospel* domina uma direção específica).

Visualização 3D com PCA com 15 componentes e Análise de K=3



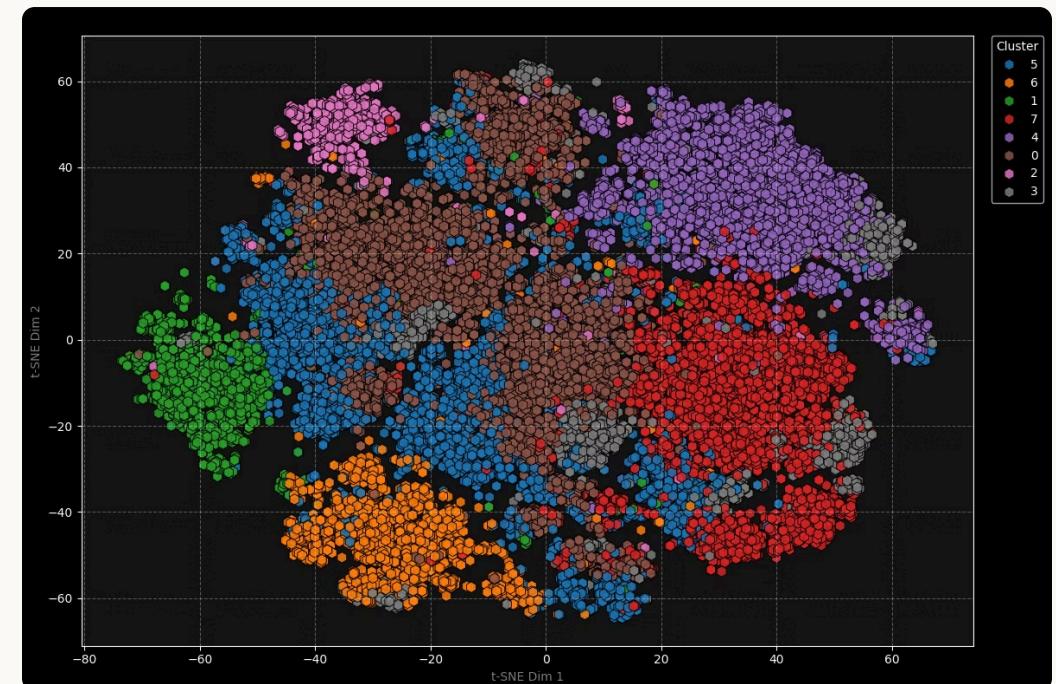
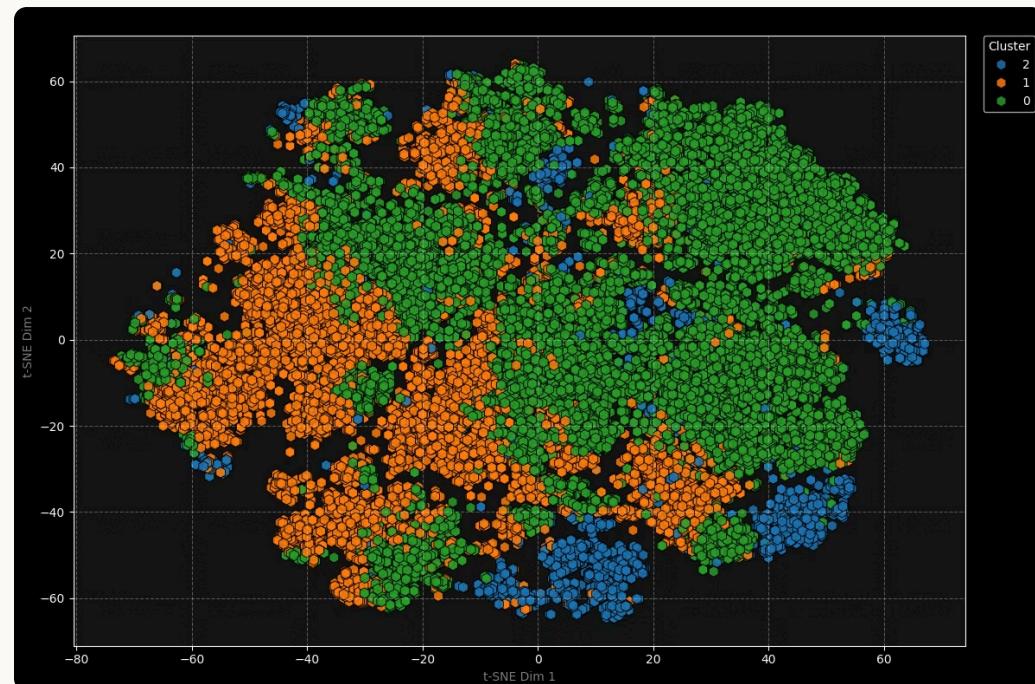
Visualização 3D com PCA com 15 componentes e Análise de K=8



Visualização de Alta Dimensionalidade: t-SNE

t-SNE (t-Distributed Stochastic Neighbor Embedding)

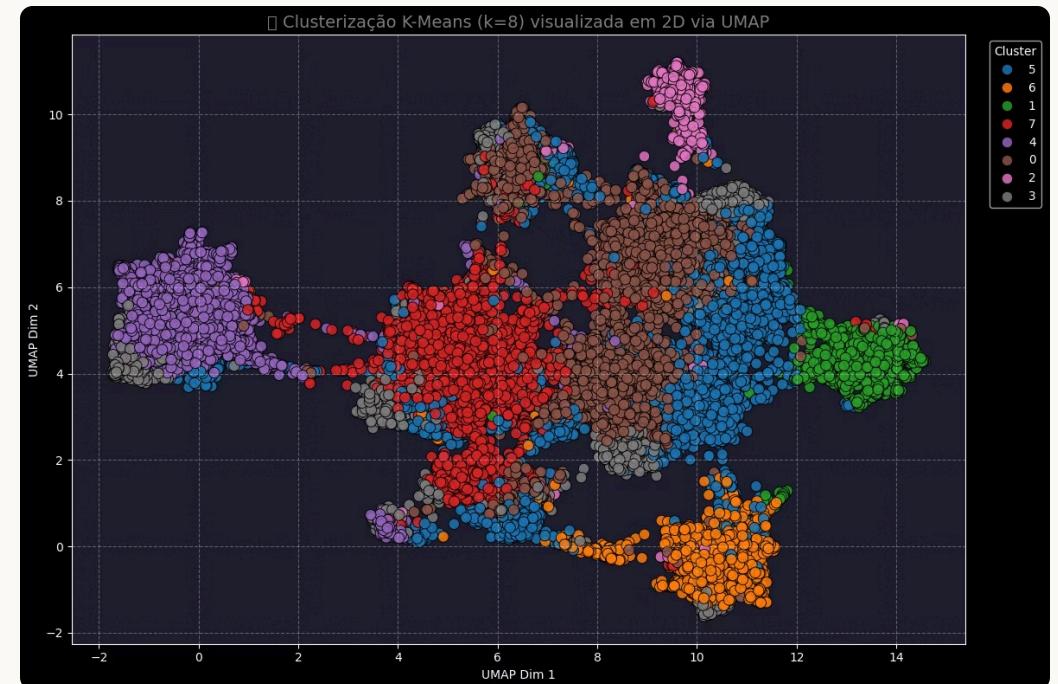
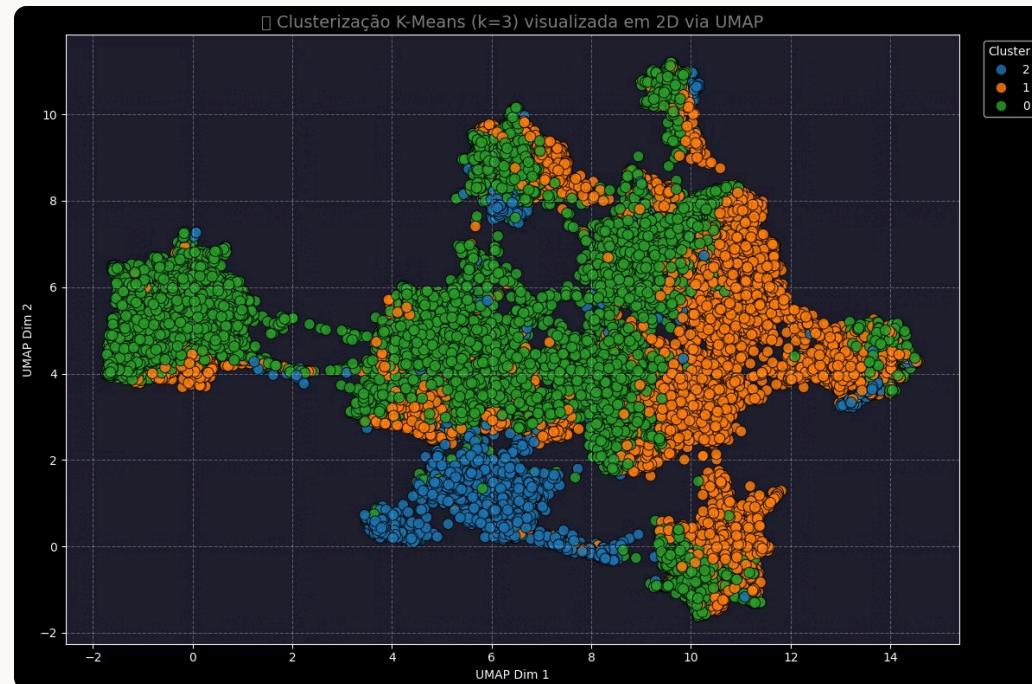
É uma técnica de redução de dimensionalidade não linear, ideal para explorar dados complexos e de alta dimensão. Focamos na visualização dos clusters resultantes do K-Means com t-SNE para k=3 e k=8, revelando a separabilidade dos grupos.



Visualização de Alta Dimensionalidade: UMAP

UMAP (Uniform Manifold Approximation and Projection)

É uma técnica de redução de dimensionalidade não linear que projeta os dados de alta dimensão para 2D ou 3D, preservando tanto a estrutura local quanto a global. Ele é especialmente eficaz para dados com padrões complexos, como músicas com múltiplas características.



Interpretação dos Clusters (k=3)

Com base nas estatísticas descritivas de cada cluster, identificamos três grupos principais de músicas, ilustrados a seguir:



Cluster 0: Músicas Energéticas e Positivas

Características: Alta energy, valence (positividade), danceability e loudness. Baixa acousticness e instrumentalness.

Temas Comuns: Provavelmente focadas em romantic, communication e like/girls.

Descrição: Engloba músicas pop, eletrônicas e dançantes, ideais para ambientes festivos e cheios de energia.

Cluster 1: Músicas Acústicas e Introspectivas

Características: Alta acousticness, com baixa energy, danceability e valence. Alta instrumentalness em alguns casos.

Temas Comuns: Mais relevantes para sadness, feelings e world/life.

Descrição: Agrupa baladas, músicas folk, clássicas ou instrumentais, com sonoridade mais calma e introspectiva.

Cluster 2: Músicas com Tópicos Específicos

Características: Variedade nas features de áudio, mas com médias mais altas em tópicos como family/gospel, obscene e violence.

Temas Comuns: Líricas que definem o cluster, como Hip Hop, Gospel ou letras explícitas.

Descrição: Definido mais fortemente pelos temas líricos do que pelas características de áudio, englobando gêneros com foco em conteúdo específico.

Esta é uma interpretação baseada em médias e medianas. A variação (desvio padrão, quartis) dentro de cada cluster também é crucial para entender a diversidade musical em cada grupo.

Interpretação Detalhada dos Clusters (k=8) – Parte 1

Analisando as estatísticas descritivas para cada um dos 8 clusters, podemos identificar perfis musicais mais distintos:



Cluster 0: Músicas Energéticas e Dançantes

Caracterizadas por altas médias em `energy`, `valence`, `danceability` e `loudness`, e baixas em `acousticness` e `instrumentalness`. Este cluster agrupa músicas pop, eletrônicas, hip-hop ou rock animado, ideais para dançar e ambientes festivos.

Cluster 1: Músicas Românticas e Melancólicas

Com altas médias em `romantic`, `sadness` e `feelings`, e médias moderadas em `energy` e `valence`. Este cluster reúne baladas, músicas românticas ou faixas com letras focadas em sentimentos e tristeza.

Cluster 2: Músicas com Foco em Performance/Audiência

Destacam-se pela alta média em `shake the audience`. Podem incluir músicas ao vivo, hinos, ou faixas projetadas para performance e engajamento do público, possivelmente em gêneros como rock, gospel ou pop performático.

Cluster 3: Músicas com Conteúdo Explícito/Violento

Com altas médias em `obscene` e `violence`, este cluster é dominado por gêneros como Hip Hop, Rap ou Metal, com letras que contêm linguagem explícita ou temas de violência.

Interpretação Detalhada dos Clusters (k=8) – Parte 2



Cluster 4: Músicas com Foco em Comunicação/Interação

Caracterizadas por uma alta média em `communication`. Este cluster inclui músicas com foco em diálogo, conversas ou interações sociais nas letras, possivelmente em gêneros como R&B, Pop ou Hip Hop.

Cluster 5: Músicas Acústicas e Instrumentais

Com altas médias em `acousticness` e `instrumentalness`, e baixas em `energy`, `valence`, `danceability` e `loudness`. Abrange música clássica, jazz, trilhas sonoras, música ambiente ou outros estilos predominantemente instrumentais e acústicos.

Cluster 6: Músicas com Temas de Família/Espiritualidade

Destacam-se por altas médias em `family/gospel` e `family/spiritual`. Provavelmente contêm músicas Gospel, religiosas ou faixas com temas fortes de família e espiritualidade.

Cluster 7: Músicas com Foco em "Like/Girls"

Caracterizadas por uma alta média em `like/girls`. Este cluster pode agrupar músicas pop ou R&B com letras focadas em atração, relacionamentos ou temas relacionados a "gostar de garotas".

Conclusão: Aumentar o número de clusters para 8 permitiu uma segmentação mais granular, revelando grupos distintos baseados tanto em características de áudio quanto em temas líricos. Isso oferece uma visão mais rica da diversidade musical no dataset.

Comparativo e Resumo da Análise de Clusters

Comparação entre K=3 e K=8

A transição de K=3 para K=8 permitiu uma segmentação mais granular dos dados musicais. Enquanto K=3 ofereceu agrupamentos amplos (Energético/Positivo, Acústico/Introspectivo, Tópicos Específicos), K=8 revelou perfis mais refinados, detalhando nuances de áudio e temas líricos.

Os clusters amplos de K=3 desdobraram-se em K=8, oferecendo distinções mais claras baseadas em características de áudio e temas líricos específicos (e.g., romântico, explícito, comunicação, família/espiritualidade).

Utilidade das Visualizações

-  **PCA (2D/3D)**: Útil para visualizar a estrutura global dos dados e a separação dos clusters.
-  **t-SNE & UMAP**: Essenciais para revelar a estrutura local e global, mostrando clusters compactos e bem separados.
-  **Dendrograma**: Ajuda a entender a hierarquia dos clusters e a sugerir o número ideal de K.

Em conjunto, essas visualizações complementares foram cruciais para interpretar e validar os resultados do K-Means.

Implicações e Utilidade Prática

As descobertas deste projeto têm diversas aplicações potenciais para a indústria musical e pesquisadores.



Sistemas de Recomendação Aprimorados

Agrupamentos de músicas podem aprimorar significativamente as recomendações personalizadas para usuários.



Análise de Tendências Musicais

Permite identificar a evolução de gêneros e estilos musicais ao longo das décadas.



Criação de Playlists Inteligentes

Facilita a curadoria de playlists temáticas com base em características musicais semelhantes.



Detecção de Similaridades e Plágio

Oferece a possibilidade de identificar similaridades em composições, auxiliando na proteção de direitos autorais.

Propostas de Estudos Futuros

Para trabalhos futuros, sugere-se aprofundar as seguintes áreas:



Validação Aprofundada de K

Empregar métricas quantitativas adicionais, como a estatística gap, para fornecer uma validação robusta na escolha do número ótimo de clusters.

Enriquecimento do Conjunto de Dados

Incorporar novas features, como metadados dos artistas (gênero, popularidade) e ano de lançamento, para revelar dimensões de similaridade inexploradas.

Modelagem Preditiva Avançada

Utilizar os clusters gerados como variável alvo em um modelo de classificação, permitindo prever o perfil de novas músicas com base em suas características.

Referências Bibliográficas

1. Fonte de Dados

Shahane, S. (2019). *Music Dataset: 1950 to 2019* [Dataset]. Kaggle.

Disponível em: <https://www.kaggle.com/datasets/saurabhshahane/music-dataset-1950-to-2019>

Bibliotecas e Ferramentas (com Documentação Oficial)

1. Python Libraries

- pandas development team. (2023). *pandas: Powerful data structures for data analysis* [Software].
Disponível em: <https://pandas.pydata.org>
- Matplotlib Development Team. (2023). *Matplotlib: Visualization with Python* [Software].
Disponível em: <https://matplotlib.org>
- scikit-learn developers. (2023). *scikit-learn: Machine Learning in Python* [Software].
Disponível em: <https://scikit-learn.org>
- McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection* [Software].
Disponível em: <https://umap-learn.readthedocs.io>
- Kaggle. (2023). *kagglehub: Kaggle Dataset Access Library* [Software].
Disponível em: <https://github.com/Kaggle/kagglehub>