Análise de Popularidade das Músicas no Spotify

Projeto Aplicado I
Turma 201825166.000.02A

Bianca Rezk de Angelo Correa
RA - 10435117
Bruna Fagundes Pereira Queiroz
RA - 10433417
Caio Cesar Teixeira Rocha
RA - 10435091

Universidade Presbiteriana Mackenzie
Faculdade de Computação e Informática
Tecnologia em Ciência de Dados

26 de novembro de 2024

Resumo

Este trabalho apresenta uma análise detalhada da popularidade das músicas na plataforma Spotify, utilizando dados extraídos da API pública da empresa. O estudo examina diversos atributos musicais, como energia, dançabilidade e valência, para identificar padrões e fatores que contribuem para o sucesso de uma música na plataforma. Através de técnicas de análise de dados e machine learning, buscamos fornecer insights valiosos para a indústria musical e para o próprio Spotify, visando aprimorar a experiência do usuário e otimizar estratégias de recomendação musical.

Sumário

Lista de Figuras

1 Introdução

O streaming musical revolucionou a forma como consumimos música, e o Spotify se destaca como uma das plataformas líderes nesse mercado. Fundado em 2006, o Spotify não apenas oferece acesso a milhões de músicas e podcasts, mas também utiliza tecnologia avançada e análise de dados para personalizar a experiência do usuário.

A análise de dados no setor de streaming musical é crucial por várias razões:

- Permite entender melhor as preferências dos usuários
- Ajuda na criação de algoritmos de recomendação mais eficientes
- Fornece insights valiosos para artistas e produtores musicais
- Auxilia na identificação de tendências musicais emergentes

Este estudo se propõe a examinar os fatores que influenciam a popularidade das músicas no Spotify, utilizando um conjunto de dados rico em informações sobre características musicais e métricas de popularidade.

2 Objetivos

Os objetivos específicos deste estudo são:

1. Identificar padrões de popularidade entre as músicas no Spotify
2. Analisar a relação entre atributos musicais (como energia, dançabilidade, valência) e a popularidade das faixas
3. Desenvolver um modelo preditivo para estimar a popularidade potencial de novas músicas
4. Fornecer insights acionáveis para artistas e produtores musicais

3 Organização da Empresa

- **Nome da empresa:** Spotify.
- **Missão/Visão/Valores:**
  - **Missão:** A missão do Spotify é revelar o poder da criatividade humana, dando a um milhão de artistas criativos a oportunidade de viver de sua arte e a bilhões de fãs a oportunidade de curtir e se inspirar nela.
  - **Visão:** Tornar-se o maior provedor de áudio global, permitindo o acesso fácil e democrático à música.
  - **Valores:** Inovação, transparência, colaboração, foco no usuário.
- **Segmento de Atuação:**
  - **Segmento Principal:** Streaming de música.

  - **Outros Segmentos:** Produção de conteúdo original, como podcasts e parcerias com artistas para lançamentos exclusivos.

- **Market Share e Posicionamento:** No segundo trimestre de 2022, o Spotify teve uma participação global de 30,5% no mercado de streaming, enquanto a Apple Music ficou em segundo lugar com 13,8%.

- **Número de Colaboradores:** Mais de 8.000 colaboradores em todo o mundo.

3.1 Iniciativas na Área de Data Science

O Spotify utiliza ciência de dados para personalizar a experiência dos usuários e otimizar a entrega de conteúdo musical. A seguir, são apresentadas algumas das principais iniciativas da empresa nesta área:

- **Recomendações Personalizadas (Discover Weekly e Release Radar)**
  - **Discover Weekly:** Essa playlist é gerada automaticamente todas as semanas e oferece aos usuários músicas que eles ainda não ouviram, mas que provavelmente irão gostar. A recomendação é feita com base em algoritmos que analisam o histórico de escuta do usuário e o comportamento de usuários com gostos semelhantes.
  - **Release Radar:** Essa playlist é atualizada toda sexta-feira e apresenta lançamentos de artistas que o usuário segue ou que o Spotify acredita serem de interesse, com base no histórico de escuta.

- **Spotify Wrapped:**
  - Um dos projetos mais populares do Spotify, o *Spotify Wrapped* é um resumo anual da atividade do usuário, incluindo suas músicas mais ouvidas, gêneros favoritos e muito mais. Ele utiliza análise de dados para criar gráficos e relatórios personalizados para cada usuário, gerando um grande engajamento nas redes sociais.

- **Taste Profiles:**
  - O Spotify utiliza um modelo de análise chamado *Taste Profiles*, que analisa o comportamento do usuário em termos de músicas, artistas e gêneros preferidos. Isso permite que a plataforma crie playlists mais precisas e personalize a experiência de descoberta de músicas.

- **Algoritmos de Skip Prediction:**
  - O Spotify desenvolveu algoritmos para prever quando um usuário pode pular uma música. Com base em dados de comportamento, como a frequência com que uma música é pulada, a empresa pode ajustar as recomendações e otimizar a experiência de audição.

- **Mood and Contextual Playlists:**

  - O Spotify também utiliza análise de dados para criar playlists contextuais, como “Songs to Sing in the Shower” ou “Workout Mix”, que são sugeridas com base na hora do dia, clima e comportamento anterior do usuário. Isso é feito com base em modelos que analisam padrões de uso e contexto.

- **Trabalhos em Destaque:**
  - Desenvolvimento de algoritmos de recomendação avançados para sugerir músicas e playlists personalizadas.
  - Parcerias com artistas para promover novos lançamentos e exclusividades.

4 Apresentação dos Dados

4.1 Nome do Conjunto de Dados

Top Spotify Tracks of 2017

4.2 Descrição

Este conjunto de dados contém informações detalhadas sobre as músicas mais populares no Spotify em 2017, incluindo atributos como popularidade, duração, energia, dançabilidade, entre outros. O objetivo é analisar quais características contribuem para o sucesso de uma música na plataforma.

4.3 Fonte

Os dados foram obtidos do Kaggle, um repositório online de conjuntos de dados para análise. O conjunto específico utilizado neste estudo é "Top Tracks of 2017"[2].

4.4 Formato

CSV (Comma-Separated Values)

5 Metadados

Os metadados das músicas fornecem uma compreensão detalhada das características sonoras e estruturais presentes no dataset, abrangendo desde aspectos gerais como o nome e artista da música até propriedades musicais específicas. A seguir, detalhamos os principais atributos, incluindo seus tipos de dados:

- **id (String):** Identificador único do Spotify para cada faixa.
- **name (String):** Nome da música no Spotify.
- **artists (String):** Artista(s) responsável(eis) pela criação da faixa.
- **danceability (Float):** Mede a adequação de uma música para dançar, com base em elementos como ritmo, estabilidade e força da batida. O valor varia de 0,0 (menos dançável) a 1,0 (mais dançável).

- **energy (Float):** Avalia a intensidade e a atividade percebida em uma faixa. Músicas com alta energia são rápidas, altas e dinâmicas, enquanto valores baixos indicam uma atmosfera mais calma. A métrica varia de 0,0 a 1,0.
- **key (Inteiro):** Representa a tonalidade da música utilizando a notação padrão de Classe de Pitches, onde cada número corresponde a uma nota específica.
- **loudness (Float):** Intensidade geral de uma faixa medida em decibéis (dB), variando tipicamente entre -60 e 0 dB. A intensidade é calculada como uma média ao longo de toda a faixa.
- **mode (Inteiro):** Indica a modalidade da faixa (maior ou menor), onde 1 representa modo maior e 0 modo menor, refletindo a escala usada para a composição melódica.
- **speechiness (Float):** Quantifica a presença de elementos falados em uma faixa. Valores altos (próximos a 1,0) indicam que a faixa possui características predominantemente de fala, enquanto valores baixos representam faixas com pouca ou nenhuma fala.
- **acousticness (Float):** Mede a probabilidade de uma faixa ser acústica, com valores variando de 0,0 (não acústica) a 1,0 (altamente acústica).
- **instrumentalness (Float):** Indica a presença de vocais em uma faixa. Valores próximos de 1,0 sugerem uma alta probabilidade de que a faixa seja instrumental, sem vocais predominantes.
- **liveness (Float):** Detecta a presença de uma audiência na gravação. Valores altos (acima de 0,8) indicam que a faixa provavelmente foi gravada ao vivo.
- **valence (Float):** Descreve o caráter emocional transmitido por uma faixa, variando de 0,0 (mais negativa) a 1,0 (mais positiva). Faixas com valência alta tendem a soar mais felizes e alegres, enquanto valência baixa transmite emoções como tristeza ou raiva.
- **tempo (Float):** Estimativa do ritmo geral da faixa em batidas por minuto (BPM). Reflete a velocidade da música e contribui para a sua classificação rítmica.
- **duration_ms (Inteiro):** Duração total da faixa em milissegundos, permitindo a análise do comprimento de cada música.
- **time_signature (Inteiro):** Estimativa da métrica musical, indicando o número de batidas por compasso. Utilizada para categorizar o padrão rítmico da música.

6 Repositório para Versionamento de Código

- `https://github.com/brunaffagundes/Data_Analysis_Project/tree/main`

O versionamento de código é fundamental para cientista de dados por diversos motivos:

- Controle de Histórico
- Colaboração

- Reprodutibilidade
- Documentação e Comentários
- Segurança e Backup
- Ciclo de Desenvolvimento

7 Cronograma

Etapa	Descrição	Data de Entrega
1	Planejamento e Definição de Objetivos	02/09/2024
2	Definição do Produto e Análise Exploratória de Dados	30/09/2024
3	Storytelling e Apresentação de Resultados	28/10/2024
4	Encerramento e Ajustes Finais	25/11/2024

8 Metodologia

Para alcançar os objetivos propostos, utilizaremos uma abordagem mista de análise exploratória de dados e modelagem preditiva. As etapas principais são:

1. **Pré-processamento dos dados:** Limpeza, tratamento de valores ausentes e normalização.
2. **Análise exploratória:** Utilização de técnicas estatísticas descritivas e visualizações para identificar padrões e correlações.
3. **Análise de correlação:** Investigação da relação entre os atributos musicais e a popularidade.
4. **Modelagem preditiva:** Desenvolvimento de modelos de machine learning (por exemplo, regressão, árvores de decisão, redes neurais) para prever a popularidade das músicas.
5. **Validação e teste:** Avaliação do desempenho dos modelos usando métricas apropriadas.
6. **Interpretação dos resultados:** Análise dos insights obtidos e suas implicações práticas.

Ferramentas a serem utilizadas incluem Python para análise de dados (com bibliotecas como pandas, numpy, scikit-learn) e ferramentas de visualização como Matplotlib e Seaborn.

9 Pensamento Computacional

- **Problema que pode ser decomposto:** A popularidade de músicas pode ser dividida em fatores como gênero, data de lançamento, playlists e localização geográfica.
- **Padrões observáveis:** Frequência de streams em relação ao tempo e ao tipo de playlist.
- **Abstração:** Generalizar os padrões para entender as preferências dos usuários em diferentes regiões.

10 Problema do Estudo

Perguntas a serem respondidas:

1. O que faz uma música ser popular?
2. Qual gênero é mais popular?
3. Qual a duração média das músicas populares?
4. Qual a frequência/bpm das músicas populares?
5. Tem algum artista específico que faz mais sucesso?
6. As músicas mais populares têm uma duração média?
7. Artistas com mais seguidores são mais populares e têm suas músicas no top 50 de forma recorrente?

11 Proposta Analítica

11.1 Análise Exploratória de Dados

- **Perguntas a serem respondidas:**
  - Quantas músicas e artistas estão presentes no dataset?
  - Quais são os gêneros musicais mais comuns?
  - Quais métricas acústicas têm maior correlação com a popularidade?
  - Existe algum outlier que possa distorcer os resultados?
  - Como as características das músicas variam entre diferentes gêneros?
  - Existe uma relação entre a duração da música e sua popularidade?
- **Número de Exemplares e Dimensões:** Quantificar e analisar o número de músicas, artistas, álbuns e gêneros no dataset.
- **Tipos de Dados:** Identificar e categorizar os tipos de dados presentes (texto, numérico, categórico, datetime).

- **Medidas de Posição e Dispersão:** Calcular e interpretar médias, medianas, desvios padrão e variâncias da popularidade, duração e outras métricas relevantes das músicas.
- **Distribuição e Frequência:** Analisar a distribuição de músicas por gênero e artista
- **Correlação:** Investigar as correlações entre os dados.
- **Valores Perdidos ou Incorretos:** Identificar, quantificar e propor estratégias para lidar com dados ausentes ou inconsistentes.
- **Anomalias e Outliers:** Detectar e analisar músicas com características excepcionais, como popularidade extrema ou métricas acústicas incomuns.

11.2 Proposta de Desenvolvimento

- **Preparação dos Dados:**
  - Limpeza e pré-processamento do dataset
  - Tratamento de valores ausentes e outliers
  - Normalização e padronização de variáveis numéricas
  - Codificação de variáveis categóricas
- **Análise Estatística:**
  - Testes de hipóteses para validar suposições sobre os dados
- **Visualização de Dados:**
  - Criação de dashboards interativos para exploração dos dados
  - Desenvolvimento de gráficos e infográficos para comunicar resultados-chave

11.3 Pipeline de Dados

Nota: O pipeline de dados descrito abaixo representa uma abordagem abrangente que seria implementada em um contexto empresarial completo. Dada a natureza deste projeto universitário, nem todas as etapas serão executadas em sua totalidade. No entanto, a descrição completa é fornecida para demonstrar o entendimento do processo e as possibilidades em um cenário real de análise de dados em larga escala.

- **Coleta de Dados:**
  - Integração com APIs de streaming de música para dados em tempo real
  - Web scraping de informações adicionais sobre artistas e álbuns
  - Implementação de sistema de coleta contínua para manter o dataset atualizado
- **Armazenamento:**
  - Configuração de um data lake para armazenamento de dados brutos

  - Implementação de um data warehouse para dados processados e prontos para análise

- **Processamento:**
  - Desenvolvimento de scripts ETL (Extract, Transform, Load) para processamento contínuo

- **Análise:**
  - Implementação de notebooks Jupyter para análise exploratória
  - Integração com ferramentas de BI para geração de relatórios automatizados

12 Análise dos Resultados

Nesta seção, apresentamos a análise dos resultados obtidos com base nas características das músicas mais populares do Spotify em 2017. Foram consideradas variáveis quantitativas como *danceability*, *energy*, *acousticness*, *duration_ms*, entre outras. A análise focou em aspectos descritivos, como médias, desvio padrão, presença de valores ausentes (NAs) e a detecção de outliers. Com 100 linhas e 16 colunas, o conjunto de dados escolhido pode ser considerado pequeno.

12.1 Estatísticas gerais dos dados

A tabela abaixo apresenta as estatísticas descritas para as variáveis numéricas presentes no conjunto de dados analisado.

	Valores Faltantes	Outliers	Valor Mínimo	Valor Máximo	Média	Variância	Desvio Padrão	Assimetria	Curtose
acousticness	0	6,0	0,000259	0,695	0,166306	2,779873e-02	0,166730	1,328141	1,108309
danceability	0	3,0	0,258000	0,927	0,696820	1,564494e-02	0,125080	-0,892722	1,515987
duration_ms	0	5,0	165387,000000	343150,000	218387,280000	1,079193e+09	32851,077720	1,565299	3,793168
energy	0	0,0	0,346000	0,932	0,660690	1,937864e-02	0,139207	-0,333296	-0,832833
instrumentalness	0	16,0	0,000000	0,210	0,004796	6,779951e-04	0,026038	6,545084	45,386667
key	0	0,0	0,000000	11,000	5,570000	1,392434e+01	3,731534	-0,094577	-1,341949
liveness	0	8,0	0,042400	0,440	0,150607	6,242704e-03	0,079011	1,398481	1,709300
loudness	0	3,0	-11,462000	-2,396	-5,652650	3,247445e+00	1,802067	-0,875746	1,150586
mode	0	0,0	0,000000	1,000	0,580000	2,460606e-01	0,496045	-0,329134	-1,930697
speechiness	0	11,0	0,023200	0,431	0,103969	9,046837e-03	0,095115	2,003411	3,511615
tempo	0	1,0	75,016000	199,864	119,202460	7,813662e+02	27,952928	0,881365	0,213949
time_signature	0	1,0	3,000000	4,000	3,990000	1,000000e-02	0,100000	-10,000000	100,000000
valence	0	0,0	0,086200	0,966	0,517049	4,684457e-02	0,216436	0,040421	-0,659314

Tabela 1: Resumo estatístico das características musicais

12.2 Estatísticas Descritivas das Variáveis

As estatísticas descritivas fornecem uma visão geral das características das músicas presentes no dataset. A seguir, destacamos os pontos principais:

- **danceability (Dançabilidade):** A média da dançabilidade foi de 0.696, indicando que as faixas têm uma tendência moderadamente alta para serem apropriadas para dançar. O desvio padrão foi de 0.125, sugerindo uma variação relativamente baixa nessa métrica entre as músicas analisadas. Foram identificados 3 outliers nesta variável, indicando que algumas músicas possuem características atípicas para dançabilidade (valores muito altos ou baixos).

- **energy (Energia):** A média de energia foi de 0.661, com um desvio padrão de 0.139, indicando que a maioria das faixas são energeticamente intensas, mas ainda existe uma variação considerável. Não foram identificados outliers, sugerindo uma distribuição relativamente consistente dessa métrica entre as músicas.

- **acousticness (Acústica):** A média da acústica foi de 0.166, com um desvio padrão de 0.166. Isso indica que a maioria das músicas no dataset não são predominantemente acústicas. Foram identificados 6 outliers, que podem corresponder a faixas que possuem uma sonoridade acústica mais destacada que o usual.

- **duration_ms (Duração em milissegundos):** A média da duração das faixas foi de 218387 ms (aproximadamente 3 minutos e 38 segundos), com um desvio padrão de 32851 ms. Foram identificados 5 outliers, indicando que algumas faixas possuem durações significativamente mais longas ou curtas.

12.3 Diversidade de Artistas no Dataset

A análise revelou a presença de 78 artistas distintos nas 100 músicas mais populares de 2017. Este número reflete uma diversidade significativa, o que sugere que a popularidade das músicas no Spotify não está concentrada em apenas alguns artistas, mas sim distribuída entre uma ampla gama de talentos. A variedade de estilos e gêneros abarcados por esses artistas demonstra o ecletismo dos ouvintes na plataforma e a capacidade do Spotify de atrair diferentes perfis de usuários.

12.4 Artistas com Mais Músicas no Dataset

Os artistas com o maior número de músicas no dataset foram Ed Sheeran e The Chainsmokers, ambos com 4 faixas, seguidos por Drake e Martin Garrix, cada um com 3 faixas. Esses artistas dominantes representam diferentes gêneros musicais, como pop e EDM, o que pode indicar uma forte preferência dos ouvintes por esses estilos em 2017. A presença significativa de alguns artistas no ranking das músicas mais populares também sugere uma maior produção musical e um maior número de lançamentos por parte deles durante esse período.

12.5 Análise da Duração das Músicas

A duração média das músicas foi de aproximadamente 3 minutos e 38 segundos, o que está dentro do padrão de tempo comumente observado para músicas populares. Faixas com essa duração equilibram bem a retenção de atenção do ouvinte e a otimização para consumo em plataformas de streaming. Músicas muito curtas (menos de 3 minutos) ou muito longas (acima de 4 minutos) tendem a se desviar do padrão ideal para streaming, sendo menos frequentes entre as mais populares.

12.6 Perfil Sonoro das Músicas Populares

Com valores médios de *danceability* e *energy* superiores a 0.6, as músicas populares de 2017 no Spotify tendem a ser altamente dançantes e enérgicas. Essa característica reflete a preferência do público por músicas que possam ser utilizadas em contextos sociais e de entretenimento, como festas e academias. Por outro lado, o baixo valor médio de

acousticness (0.166) indica que as músicas populares eram predominantemente produzidas de forma eletrônica, com poucos elementos acústicos, o que está alinhado com a tendência crescente de música eletrônica e hip-hop nos charts globais durante aquele ano.

12.7 Análise de Valores Ausentes

Todas as colunas analisadas não apresentaram valores ausentes (NAs), o que sugere que o conjunto de dados está completo e não requer tratamento adicional para ausência de dados.

12.8 Detecção de Outliers

A detecção de outliers foi feita utilizando o método do intervalo interquartil (IQR). Os resultados indicaram a presença de outliers principalmente nas variáveis *acousticness*, *danceability* e *duration_ ms*. Esses outliers representam músicas com características sonoras ou durações atípicas em relação à maioria das músicas no conjunto de dados.

- **Acousticness:** Os 6 outliers identificados podem indicar músicas predominantemente acústicas, o que é raro no conjunto de faixas mais populares.
- **Danceability:** Os 3 outliers indicam músicas que se destacam por serem muito fáceis ou muito difíceis de dançar, em comparação com as demais.
- **Duration:** Os 5 outliers correspondem a músicas com duração muito superior ou inferior à média observada.

12.9 Análise de Correlações

Nesta análise, exploramos as correlações entre as variáveis do conjunto de dados das músicas mais populares no Spotify em 2017. Utilizamos um mapa de calor para visualizar as correlações e gráficos de dispersão (*scatterplots*) para explorar as relações entre variáveis selecionadas.

12.9.1 Mapa de Calor da Matriz de Correlação

A Figura 1 apresenta um **mapa de calor** que exibe a matriz de correlação das variáveis numéricas. Nele, podemos observar a força das relações entre diferentes variáveis:

- As variáveis *danceability* e *valence* apresentam uma correlação positiva moderada, sugerindo que faixas com maior dançabilidade tendem a transmitir emoções mais positivas.
- *Energy* apresenta uma correlação significativa com *loudness*, o que faz sentido, uma vez que músicas energéticas tendem a ser mais altas ou barulhentas.
- *Duration_ ms* e *tempo* não apresentam correlações significativas com outras variáveis no dataset, indicando que a duração e o tempo das faixas não influenciam fortemente outras características musicais.

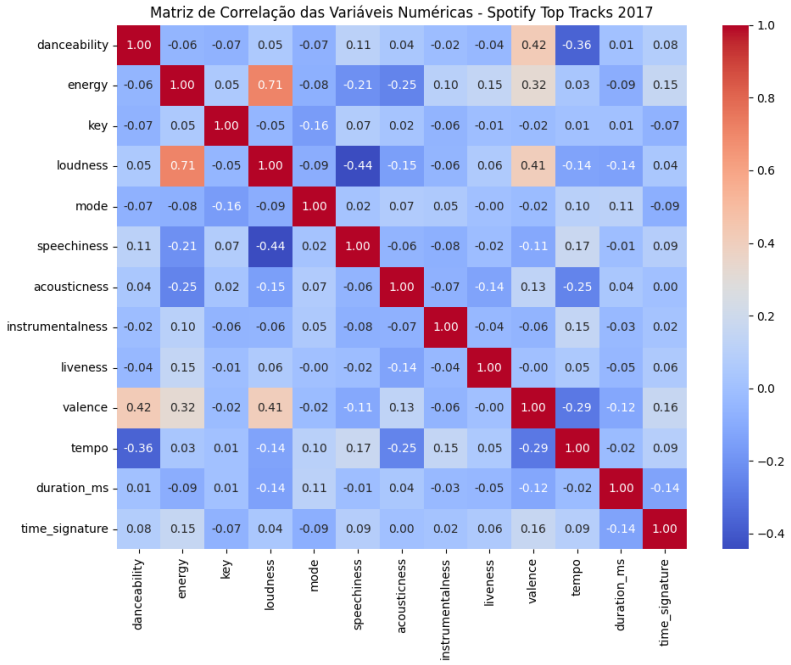

Figura 1: Matriz de Correlação das Variáveis Numéricas - Spotify Top Tracks 2017

12.9.2 Scatterplots para Variáveis Selecionadas

Além do mapa de calor, criamos scatterplots para algumas variáveis selecionadas, a fim de explorar visualmente as relações:

- ***danceability* x *energy*:** Mostra como a dançabilidade e a energia das faixas se relacionam. Podemos observar que faixas com alta energia não necessariamente apresentam maior dançabilidade, indicando que músicas intensas nem sempre são as mais adequadas para dançar.
- ***acousticness* x *valence*:** Exibe a relação entre acústica e positividade musical. Músicas mais acústicas parecem não se concentrar em uma faixa específica de *valence*, sugerindo que acústica não é um indicador forte de positividade ou negatividade emocional.
- ***duration_ms* x *danceability*:** Mostra que faixas mais curtas tendem a ter uma ligeira variação na dançabilidade, mas a duração não parece influenciar fortemente essa característica.

A Figura 2 mostra os gráficos de dispersão para as variáveis selecionadas.

12.9.3 Análise Específica de Correlação com Popularidade

Caso tivéssemos uma variável como *popularity*, seria interessante analisar a correlação dessa variável com as características das músicas, como *danceability*, *energy* e *valence*. Isso nos ajudaria a entender quais fatores mais influenciam a popularidade de uma faixa.

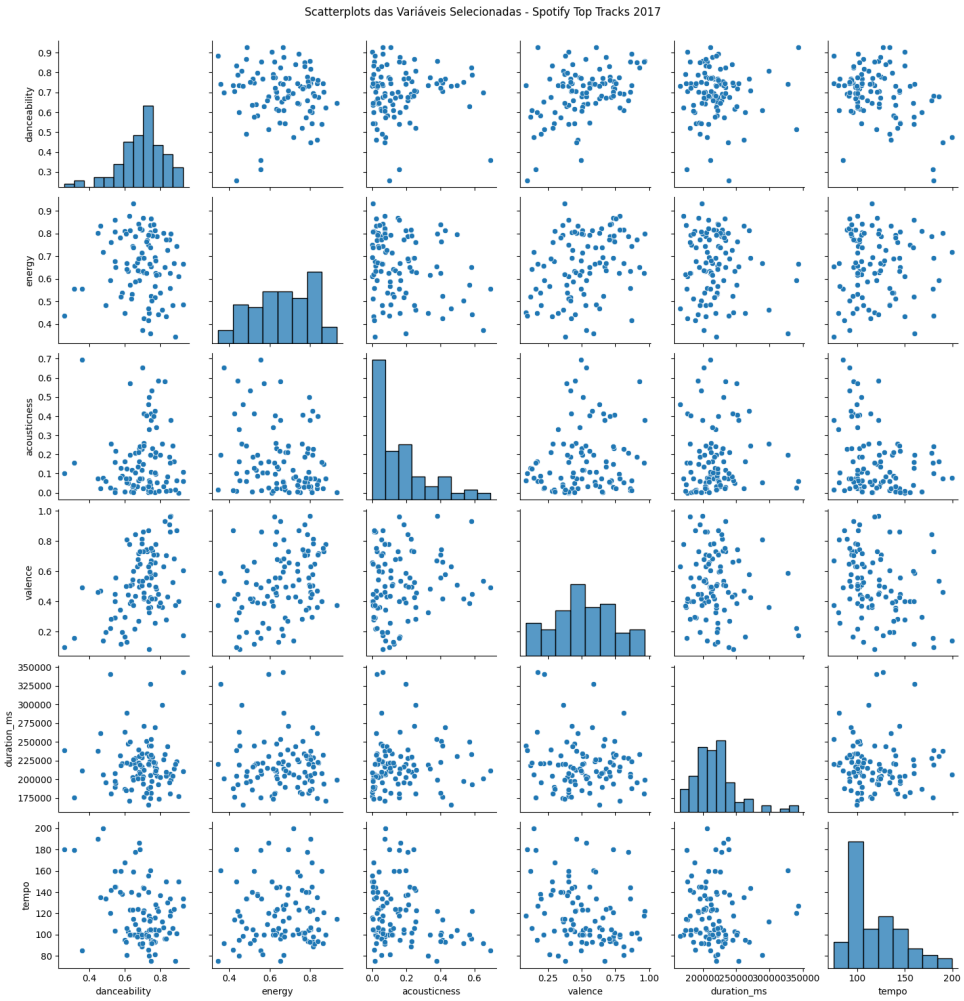

Figura 2: Scatterplots das Variáveis Selecionadas - Spotify Top Tracks 2017

12.10 Discussão Geral

Os resultados indicam que a maioria das faixas mais populares do Spotify em 2017 seguem um padrão relativamente homogêneo em termos de *danceability*, *energy* e *duration*. Observa-se que variáveis como *danceability* e *valence* têm uma correlação positiva moderada (≈ 0.5), sugerindo que faixas com maior dançabilidade tendem a transmitir emoções mais positivas. Isso significa que músicas que são mais fáceis de dançar também são percebidas como mais alegres ou animadas, o que pode influenciar sua popularidade.

Por outro lado, *energy* apresenta uma correlação negativa com *acousticness*, indicando que músicas mais energéticas tendem a ser menos acústicas, ou seja, faixas que possuem alta intensidade, ritmo rápido e maior volume não costumam ter elementos acústicos predominantes. Esse comportamento sugere que músicas com perfil acústico são geralmente mais calmas e suaves, diferenciando-se das faixas mais energéticas.

Os **scatterplots** reforçam a observação de que características como *danceability* e *energy* não possuem uma relação linear evidente, indicando que músicas intensas nem sempre são as mais dançantes. Além disso, a análise entre *duration_ms* e *danceability* mostra que a duração da faixa não influencia fortemente o quanto ela é considerada dançável, sugerindo que o tempo total de uma música não é um fator determinante para essa característica.

A presença de outliers em variáveis como *acousticness* e *danceability* mostra que há faixas no dataset que fogem do padrão comum observado. Essas músicas podem ser mais acústicas ou mais (ou menos) dançantes do que a maioria, refletindo uma diversidade de características musicais no conjunto das faixas mais populares. Essa variação pode indicar uma preferência geral dos usuários por faixas que seguem padrões similares, mas também a aceitação de músicas que fogem das normas esperadas, mostrando a diversidade e o gosto eclético do público.

Os gráficos de dispersão do **pairplot** também permitem identificar como múltiplas variáveis se relacionam entre si. Por exemplo, podemos observar que a *danceability* tende a ter uma relação mais distribuída com outras variáveis, enquanto *acousticness* e *energy* mostram padrões de distribuição opostos, reforçando a interpretação de que músicas mais acústicas são geralmente menos energéticas. Esses insights ajudam a compreender as características subjacentes das músicas mais populares e como elas se posicionam em termos de composição musical.

Em suma, as correlações e visualizações sugerem que, embora existam padrões predominantes entre algumas características, como a relação positiva entre *danceability* e *valence*, há uma diversidade considerável no conjunto de dados, refletindo a variedade de preferências musicais dos ouvintes do Spotify.

13 Conclusão

Este estudo tem como objetivo realizar uma análise inicial dos fatores que podem influenciar a popularidade das músicas no Spotify. Por meio de uma combinação de técnicas básicas de análise exploratória de dados e modelagem preditiva, esperamos obter alguns insights preliminares que possam ser úteis para a indústria musical, artistas e a própria plataforma Spotify. Os resultados deste trabalho podem ajudar a entender melhor as preferências musicais dos usuários e contribuir para a elaboração de estratégias mais simples de produção e promoção musical no ambiente de streaming.

14 Esboço do Storytelling

Introdução e Apresentação do Grupo

O vídeo começará com uma breve apresentação dos membros do grupo: Bianca, Bruna e Caio. Cada um de nós aparecerá na tela, dizendo nosso nome e curso, estabelecendo uma conexão pessoal com o público e demonstrando entusiasmo pelo projeto.

Nome do Projeto e Contextualização

Apresentaremos o título do projeto: **Análise de Popularidade das Músicas no Spotify**. Utilizaremos animações e elementos visuais relacionados à música e ao Spotify para captar a atenção do espectador desde o início.

Empresa Estudada: Spotify

Introduziremos o Spotify como a plataforma líder em streaming musical que revolucionou a forma como consumimos música. Destacaremos sua missão de conectar artistas e fãs e a importância da plataforma no cenário musical atual.

Área do Problema

Exploraremos a questão central: **O que faz uma música se tornar popular no Spotify?** Levantaremos a curiosidade do público ao mencionar que, com milhões de músicas disponíveis, entender os fatores que influenciam a popularidade é um desafio complexo e fascinante.

Descrição do Problema / Gap

Explicaremos que artistas e produtores buscam compreender quais elementos musicais contribuem para o sucesso de uma faixa. Há uma lacuna na compreensão aprofundada de como características como energia, dançabilidade e valência impactam a popularidade das músicas.

Proposta Analítica

Apresentaremos nossa abordagem analítica para responder a essa questão. Mencionaremos que utilizaremos técnicas de ciência de dados para analisar um conjunto de dados das músicas mais populares de 2017 no Spotify, buscando identificar padrões e relações entre os atributos musicais e a popularidade.

Dados Disponíveis

Mostraremos o conjunto de dados utilizado: **Top Spotify Tracks of 2017**. Destacaremos que este dataset contém informações detalhadas sobre 100 músicas, incluindo métricas como dançabilidade, energia, duração e outras características acústicas.

Análise Exploratória

Compartilharemos alguns insights da análise exploratória de dados. Por exemplo:

- Músicas mais populares tendem a ter alta dançabilidade e energia.
- A duração média das músicas populares é de aproximadamente 3 minutos e 38 segundos.
- Artistas como Ed Sheeran e The Chainsmokers tiveram maior presença nas paradas em 2017.

Utilizaremos gráficos e visualizações para tornar a apresentação mais dinâmica e visualmente atraente.

Resultados Pretendidos

Concluiremos apresentando os resultados esperados:

- Identificar quais características musicais mais influenciam a popularidade.
- Fornecer insights úteis para artistas e produtores na criação de músicas que ressoem com o público.

Encerraremos com uma mensagem convidativa, incentivando o público a refletir sobre como a ciência de dados pode transformar a indústria musical e agradecendo pela atenção.

15 Apresentação Final

O grupo elaborou um vídeo contendo a apresentação do que foi desenvolvido ao longo da disciplina.

O vídeo pode ser acessado pelo YouTube:
`https://www.youtube.com/watch?v=ERirLiuSuXc`

16 Referências Bibliográficas

Referências

[1] SPOTIFY. Spotify API Documentation. 2021. Disponível em: `https://developer.spotify.com/documentation/web-api`. Acesso em: 2 set. 2024.

[2] TAMER, N. Top Tracks of 2017. Kaggle. 2017. Disponível em: `https://www.kaggle.com/datasets/nadintamer/top-tracks-of-2017`. Acesso em: 2 set. 2024.

17 Apêndices

17.1 Apêndice A: código de análise exploratória

```
# Selecionar apenas colunas numéricas
numerical_columns = top_2017.select_dtypes(include=['float64', 'int64'])

# Gerar uma análise descritiva para todas as
# variáveis numéricas do conjunto de dados
description = numerical_columns.describe()

# Analisar a quantidade de valores ausentes (NAs) por coluna
missing_values = top_2017.isna().sum()

# Verificar a presença de outliers utilizando o
# método do IQR (Interquartile Range)
Q1 = numerical_columns.quantile(0.25)
Q3 = numerical_columns.quantile(0.75)
IQR = Q3 - Q1
outliers = (
    (numerical_columns < (Q1 - 1.5 * IQR))
    | (numerical_columns > (Q3 + 1.5 * IQR))
    ).sum()

# Gerar uma lista com as informações coletadas
# para cada variável numérica
summary = pd.DataFrame({
    'Missing Values': missing_values,
    'Outliers': outliers,
    'Min Value': description.loc['min'],
```

```
    'Max Value': description.loc['max'],
    'Mean': description.loc['mean'],
    # Variância calculada diretamente como variância amostral
    'Variance': description.loc['std']**2,
    'Standard Deviation': description.loc['std'],
})

# Exibir a tabela resumo
summary
```